© Australia (2018) www.nature.com/nature 479

nature

BRANCHING OUT

New class of plant hormones inhibits branch formation

MOMENTOUS MEETINGS
The birth of CERN

PALAEONTOLOGY
Is Argentina the new China?

TRANSPIRATION
The pulling power of a 'synthetic tree'

NATURE JOBS
Forces on phonons

nature

www.nature.com/nature

Vol 455 | Issue no. 7210 | 11 September 2008

Turning blight into bloom

As we become an ever more urban species, cities will be a crucial front in the fight against climate change. Scientists, architects and planners must join forces to make our metropolitan future clean and sustainable.

Humanity has passed a milestone: more people live in cities than in rural areas. The current rate of urbanization is unprecedented in our history. In 1950, only 29% of people lived in cities; by 2050, 70% are projected to do so — most of them in poorer countries. Among many other issues, this rapid concentration makes cities a front line in the battles against climate change and air pollution. Confronting the challenges of rampant urbanization demands integrated, multidisciplinary approaches, and new thinking.

Take the building boom associated with the increased wealth of urban areas, and its impact on greenhouse-gas emissions. In China alone, the United Nations Environmental Programme estimates that energy demand for heating homes built over the next decade could increase by some 430 terawatt-hours, or 4% of China's total energy use in 2003. Worldwide, the energy consumed by buildings already accounts for around 45% of greenhouse-gas emissions.

Fortunately, researchers in Germany and elsewhere have already shown that they can reduce that energy consumption by 80–90%, just by overhauling obsolete building designs and using existing technologies (see *Nature* **452,** 520–523; 2008). These ultra-efficient buildings demand that planners, architects, engineers and building scientists work together from the outset, and require higher levels of expertise than conventional buildings. But such buildings are often cheaper than those built using conventional methods. Research is also needed to develop technologies, materials and energy concepts, but green building research today is fragmented and poorly funded.

Expanding cities must embrace such technologies and strategies — and not just in the developed nations. It may seem utopian to promote these innovations in emerging and developing-world megacities, many of whose inhabitants can barely afford a roof over their heads. But those countries have already shown a gift for technological fast-forwarding, for example, by leapfrogging the need for landline infrastructure to embrace mobile phones. And many poorer countries have a rich tradition of adapting buildings to local practices, environments and climates — a home-grown approach to integrated design that has been all but been lost in the West. They now have an opportunity to combine these traditional approaches with modern technologies.

Integrated thinking is also needed to mitigate urban air pollution, which is becoming a serious health and environmental risk in many regions — as shown by China's struggle to clean up Beijing's air for the Olympics. Understanding air pollution will require researchers from multiple disciplines, from atmospheric chemistry to meteorology, working over scales from street level to global (see page 142). And reducing it will require integrated policies for urban planning, transport and housing — not least to reduce the use of cars.

"Many poorer countries have a rich tradition of adapting buildings to local practices."

This unified approach is not entirely new: the Brazilian city of Curitiba has linked urban planning, transport and housing development for decades; a large majority of its nearly 2 million inhabitants now use public transport and recycle waste. But new city design and planning tools are making it considerably easier. Modern simulation software, for example, allows designers to do full structural analyses of buildings and model their energy balances in three dimensions. Likewise, geographical software allows planners to model the impact of planning scenarios on variables such as traffic and emissions.

Integrated thinking and planning requires a cultural shift: most urban development has been laissez-faire. Yet there are signs that such a shift has begun. In California, state lawmakers are considering a bill to restrict transportation subsidies to compact developments that reduce the need for car travel. If passed, this would be the first legislation in the United States to tie urban planning to the control of greenhouse gases. And China plans to build one of the world's first eco-cities off the coast of Shanghai. Dongtan aims to be free of cars, energy-independent — thanks to energy-efficient buildings, plus the extensive use of renewables and geothermal strategies for heating and cooling — and home to half a million people.

Such innovations need to be encouraged, expanded and accelerated — everywhere on the planet. ■

Brave new worlds

A new series of essays looks back at scientific meetings that had world-changing consequences.

Creative ideas are not always solo strokes of genius, argues Ed Catmull, the computer-scientist president of Pixar and Disney Animation Studios, in the current issue of the *Harvard Business Review*. Frequently, he says, the best ideas emerge when talented people from different disciplines work together.

This week, *Nature* begins a series of six Essays that illustrate Catmull's case. Each recalls a conference in which a creative outcome emerged from scientists pooling ideas, expertise and time with others — especially policy-makers, non-governmental organizations and the media. Each is written by someone who was there, usually an organizer or the meeting chair. Because the conferences were chosen for their societal consequences, we've called our series 'Meetings that Changed the World'.

This week, François de Rose relives the drama of the December 1951 conference at the UNESCO headquarters in Paris that led to the creation of CERN, the European particle-physics laboratory based

nature

www.nature.com/nature

Vol 455 | Issue no. 7210 | 11 September 2008

Turning blight into bloom

As we become an ever more urban species, cities will be a crucial front in the fight against climate change. Scientists, architects and planners must join forces to make our metropolitan future clean and sustainable.

Humanity has passed a milestone: more people live in cities than in rural areas. The current rate of urbanization is unprecedented in our history. In 1950, only 29% of people lived in cities; by 2050, 70% are projected to do so — most of them in poorer countries. Among many other issues, this rapid concentration makes cities a front line in the battles against climate change and air pollution. Confronting the challenges of rampant urbanization demands integrated, multidisciplinary approaches, and new thinking.

Take the building boom associated with the increased wealth of urban areas, and its impact on greenhouse-gas emissions. In China alone, the United Nations Environmental Programme estimates that energy demand for heating homes built over the next decade could increase by some 430 terawatt-hours, or 4% of China's total energy use in 2003. Worldwide, the energy consumed by buildings already accounts for around 45% of greenhouse-gas emissions.

Fortunately, researchers in Germany and elsewhere have already shown that they can reduce that energy consumption by 80–90%, just by overhauling obsolete building designs and using existing technologies (see *Nature* **452,** 520–523; 2008). These ultra-efficient buildings demand that planners, architects, engineers and building scientists work together from the outset, and require higher levels of expertise than conventional buildings. But such buildings are often cheaper than those built using conventional methods. Research is also needed to develop technologies, materials and energy concepts, but green building research today is fragmented and poorly funded.

Expanding cities must embrace such technologies and strategies — and not just in the developed nations. It may seem utopian to promote these innovations in emerging and developing-world megacities, many of whose inhabitants can barely afford a roof over their heads. But those countries have already shown a gift for technological fast-forwarding, for example, by leapfrogging the need for landline infrastructure to embrace mobile phones. And many poorer countries have a rich tradition of adapting buildings to local practices, environments and climates — a home-grown approach to integrated design that has been all but been lost in the West. They now have an opportunity to combine these traditional approaches with modern technologies.

Integrated thinking is also needed to mitigate urban air pollution, which is becoming a serious health and environmental risk in many regions — as shown by China's struggle to clean up Beijing's air for the Olympics. Understanding air pollution will require researchers from multiple disciplines, from atmospheric chemistry to meteorology, working over scales from street level to global (see page 142). And reducing it will require integrated policies for urban planning, transport and housing — not least to reduce the use of cars.

"Many poorer countries have a rich tradition of adapting buildings to local practices."

This unified approach is not entirely new: the Brazilian city of Curitiba has linked urban planning, transport and housing development for decades; a large majority of its nearly 2 million inhabitants now use public transport and recycle waste. But new city design and planning tools are making it considerably easier. Modern simulation software, for example, allows designers to do full structural analyses of buildings and model their energy balances in three dimensions. Likewise, geographical software allows planners to model the impact of planning scenarios on variables such as traffic and emissions.

Integrated thinking and planning requires a cultural shift: most urban development has been laissez-faire. Yet there are signs that such a shift has begun. In California, state lawmakers are considering a bill to restrict transportation subsidies to compact developments that reduce the need for car travel. If passed, this would be the first legislation in the United States to tie urban planning to the control of greenhouse gases. And China plans to build one of the world's first eco-cities off the coast of Shanghai. Dongtan aims to be free of cars, energy-independent — thanks to energy-efficient buildings, plus the extensive use of renewables and geothermal strategies for heating and cooling — and home to half a million people.

Such innovations need to be encouraged, expanded and accelerated — everywhere on the planet. ■

Brave new worlds

A new series of essays looks back at scientific meetings that had world-changing consequences.

Creative ideas are not always solo strokes of genius, argues Ed Catmull, the computer-scientist president of Pixar and Disney Animation Studios, in the current issue of the *Harvard Business Review*. Frequently, he says, the best ideas emerge when talented people from different disciplines work together.

This week, *Nature* begins a series of six Essays that illustrate Catmull's case. Each recalls a conference in which a creative outcome emerged from scientists pooling ideas, expertise and time with others — especially policy-makers, non-governmental organizations and the media. Each is written by someone who was there, usually an organizer or the meeting chair. Because the conferences were chosen for their societal consequences, we've called our series 'Meetings that Changed the World'.

This week, François de Rose relives the drama of the December 1951 conference at the UNESCO headquarters in Paris that led to the creation of CERN, the European particle-physics laboratory based

near Geneva (see page 174). De Rose, then France's representative to the United Nations Atomic Energy Commission, chaired the meeting. He had got caught up in the process after becoming friends with Robert Oppenheimer, one of CERN's earliest proponents. De Rose said in a separate interview with *Nature* that CERN was the result of the capacity of scientists such as Oppenheimer to propose grand ideas, and worry about obstacles later.

Although this approach does not always work, the next few weeks will show that it really has changed the world. In the ensuing half-century, CERN has revolutionized our understanding of the subatomic world; with the switching-on this week of the Large Hadron Collider (see page 156) it promises to scale new heights.

When we began to think about commissioning this series, several difficulties arose. First, we were looking for more than the traditional scientific conference, and it was notable how few of the twentieth century's world-changing meetings had involved scientists taking a lead. As a list emerged, we were faced with another problem: time had sadly depleted the pool of writers. This week's author, for example, is among the few surviving members of a group that met 57 years ago.

The six events that made the final cut took place on three continents and span five decades, from 1951 to the dawn of the new millennium. They represent the twentieth century's promise, and two of its greatest threats. And they illustrate a period in history when scientists felt they should raise a collective voice to advance the public good. The six meetings have something else in common. In wanting to change their world, the scientists involved needed and obtained the support of governments and, in some cases, the media.

In two of the conferences — those related to CERN and the Human Genome Project — scientists organized themselves and others to create new and exciting research endeavours. But the other meetings considered in our series had very different aims. At a 1975 conference held in Asilomar, California, for example, geneticists felt compelled to sound an alarm over DNA modification, then a new technology of uncertain impact. At a meeting in Bellagio, Italy, in 1969, plant scientists were among those who convinced governments and philanthropic foundations to invest in technologies to take the green revolution to the developing world.

As in any series of this nature, some caveats are in order. First, global initiatives are a process in which many decisions are made over many years. In the case of CERN, the Paris 1951 event was not the first official meeting in the institution's history. It was, however, an occasion where private disagreements between governments became public, and where a consensus was eventually found to move the project forward. Without such a consensus, it is debatable whether CERN would have taken the direction it did.

Second, our list is not the final word. There are other candidates for the title of Meetings that Changed the World. And our illustrious attendees' opinions are, of course, personal and often provocative. Readers are invited to have their say at http://network.nature.com/forums/naturenewsandopinion/2359. ■

A bigger picture

Beneath cancer's daunting complexity lies a simplicity that gives grounds for hope.

For several years now, large-scale cancer-genome studies have made it increasingly clear that a tumour cell is a genetic disaster area littered with mutations that differ not only from one type of cancer to the next, but from one patient to the next. Pharmaceutical companies have had to accept that Gleevec, a drug that treats a form of leukaemia by targeting a specific gene product, is almost certainly going to be a rare exception in the therapeutic arsenal; most cancers are far too complex to yield to such a magic bullet.

That message was hammered home with new statistical power in three studies released last week (see page 148). Two of the studies, published in *Science* by a team based at Johns Hopkins Kimmel Cancer Center in Baltimore, Maryland, focused on pancreatic cancer and a type of brain cancer called glioblastoma multiforme — both among the most fatal and intractable tumours known. The third paper, published in *Nature* by the Cancer Genome Atlas project, also focused on glioblastoma. The studies took a more comprehensive approach than previous large cancer-genomics studies, by simultaneously analysing genetic sequences, copy-number variations, expression arrays and other forms of data. The Johns Hopkins team looked at all the active genes in tumours from a few dozen patients; the Genome Atlas team looked at selected genes in tumours from 206 patients. Taken together, their results show that no single mutated gene lies at the heart of any of these tumours. The pancreatic tumour samples, for example, showed an average of 63 genetic mutations each — with considerable variation from one sample to the next.

That conclusion might make the prospects for new targeted drug therapies for cancer seem hopeless. And yet, the reality may be just the opposite. The richness of the data becoming available in these and other studies allows researchers to cut through the complexity. Genes work together in pathways of reactions to accomplish a particular biological function, such as cell division — and many or most of the mutated genes picked up by these cancer studies are involved in a comparatively small number of pathways. The Johns Hopkins team found that most of the mutations in their pancreatic tumours affected just 12 pathways. The Genome Atlas team found that most of its glioblastomas showed mutations in a set of three pathways. So drugs targeting these pathways might work in more patients than drugs that target only one of a pathway's myriad gene components.

To realize that hope, researchers and funding agencies will need to do many more such studies on many more types of cancer. Just as important is the next step, which is to determine how these mutated pathways contribute to the development of cancer — and how that contribution might be removed. After that comes the task of finding useful biomarkers, chemical signals that will allow therapists to determine which pathways have been affected in each cancer patient, and how that patient will respond to any given therapy.

None of this will be easy. Untangling the immense complexity of cancer will be big science by anyone's definition, requiring a long-term commitment and enormous amounts of data. And yet, that very complexity has begun to give reason for optimism. ■

near Geneva (see page 174). De Rose, then France's representative to the United Nations Atomic Energy Commission, chaired the meeting. He had got caught up in the process after becoming friends with Robert Oppenheimer, one of CERN's earliest proponents. De Rose said in a separate interview with *Nature* that CERN was the result of the capacity of scientists such as Oppenheimer to propose grand ideas, and worry about obstacles later.

Although this approach does not always work, the next few weeks will show that it really has changed the world. In the ensuing half-century, CERN has revolutionized our understanding of the subatomic world; with the switching-on this week of the Large Hadron Collider (see page 156) it promises to scale new heights.

When we began to think about commissioning this series, several difficulties arose. First, we were looking for more than the traditional scientific conference, and it was notable how few of the twentieth century's world-changing meetings had involved scientists taking a lead. As a list emerged, we were faced with another problem: time had sadly depleted the pool of writers. This week's author, for example, is among the few surviving members of a group that met 57 years ago.

The six events that made the final cut took place on three continents and span five decades, from 1951 to the dawn of the new millennium. They represent the twentieth century's promise, and two of its greatest threats. And they illustrate a period in history when scientists felt they should raise a collective voice to advance the public good. The six meetings have something else in common. In wanting to change their world, the scientists involved needed and obtained the support of governments and, in some cases, the media.

In two of the conferences — those related to CERN and the Human Genome Project — scientists organized themselves and others to create new and exciting research endeavours. But the other meetings considered in our series had very different aims. At a 1975 conference held in Asilomar, California, for example, geneticists felt compelled to sound an alarm over DNA modification, then a new technology of uncertain impact. At a meeting in Bellagio, Italy, in 1969, plant scientists were among those who convinced governments and philanthropic foundations to invest in technologies to take the green revolution to the developing world.

As in any series of this nature, some caveats are in order. First, global initiatives are a process in which many decisions are made over many years. In the case of CERN, the Paris 1951 event was not the first official meeting in the institution's history. It was, however, an occasion where private disagreements between governments became public, and where a consensus was eventually found to move the project forward. Without such a consensus, it is debatable whether CERN would have taken the direction it did.

Second, our list is not the final word. There are other candidates for the title of Meetings that Changed the World. And our illustrious attendees' opinions are, of course, personal and often provocative. Readers are invited to have their say at http://network.nature.com/forums/naturenewsandopinion/2359. ■

A bigger picture

Beneath cancer's daunting complexity lies a simplicity that gives grounds for hope.

For several years now, large-scale cancer-genome studies have made it increasingly clear that a tumour cell is a genetic disaster area littered with mutations that differ not only from one type of cancer to the next, but from one patient to the next. Pharmaceutical companies have had to accept that Gleevec, a drug that treats a form of leukaemia by targeting a specific gene product, is almost certainly going to be a rare exception in the therapeutic arsenal; most cancers are far too complex to yield to such a magic bullet.

That message was hammered home with new statistical power in three studies released last week (see page 148). Two of the studies, published in *Science* by a team based at Johns Hopkins Kimmel Cancer Center in Baltimore, Maryland, focused on pancreatic cancer and a type of brain cancer called glioblastoma multiforme — both among the most fatal and intractable tumours known. The third paper, published in *Nature* by the Cancer Genome Atlas project, also focused on glioblastoma. The studies took a more comprehensive approach than previous large cancer-genomics studies, by simultaneously analysing genetic sequences, copy-number variations, expression arrays and other forms of data. The Johns Hopkins team looked at all the active genes in tumours from a few dozen patients; the Genome Atlas team looked at selected genes in tumours from 206 patients. Taken together, their results show that no single mutated gene lies at the heart of any of these tumours. The pancreatic tumour samples, for example, showed an average of 63 genetic mutations each — with considerable variation from one sample to the next.

That conclusion might make the prospects for new targeted drug therapies for cancer seem hopeless. And yet, the reality may be just the opposite. The richness of the data becoming available in these and other studies allows researchers to cut through the complexity. Genes work together in pathways of reactions to accomplish a particular biological function, such as cell division — and many or most of the mutated genes picked up by these cancer studies are involved in a comparatively small number of pathways. The Johns Hopkins team found that most of the mutations in their pancreatic tumours affected just 12 pathways. The Genome Atlas team found that most of its glioblastomas showed mutations in a set of three pathways. So drugs targeting these pathways might work in more patients than drugs that target only one of a pathway's myriad gene components.

To realize that hope, researchers and funding agencies will need to do many more such studies on many more types of cancer. Just as important is the next step, which is to determine how these mutated pathways contribute to the development of cancer — and how that contribution might be removed. After that comes the task of finding useful biomarkers, chemical signals that will allow therapists to determine which pathways have been affected in each cancer patient, and how that patient will respond to any given therapy.

None of this will be easy. Untangling the immense complexity of cancer will be big science by anyone's definition, requiring a long-term commitment and enormous amounts of data. And yet, that very complexity has begun to give reason for optimism. ■

RESEARCH HIGHLIGHTS

CLIMATE CHANGE

'Hockey stick' holds up

Proc. Natl Acad. Sci. USA **105,** 13252–13257 (2008)

A fresh analysis of climate indicators shows that the Northern Hemisphere is warmer now than it has been in at least 1,300 years.

Previous analyses of climatic history by Michael Mann of Pennsylvania State University in University Park and his colleagues produced a distinctive 'hockey stick' shape; but some of this analysis, and the tree-ring data it used, came under attack.

The latest work by Mann and his co-workers involves various climate proxies, including corals, ice cores, historical records and marine sediments. The authors show that current warming is anomalous even if all tree-ring data are eschewed.

SEXUAL IMPRINTING

Facing Oedipus

Proc. R. Soc. B doi:10.1098/rspb.2008.1021 (2008)

The suggestion that people seek mates that resemble their parents is as old as civilization. Tamas Bereczkei and his colleagues at the University of Pécs in Hungary have found new evidence linking partner choices to parental appearance.

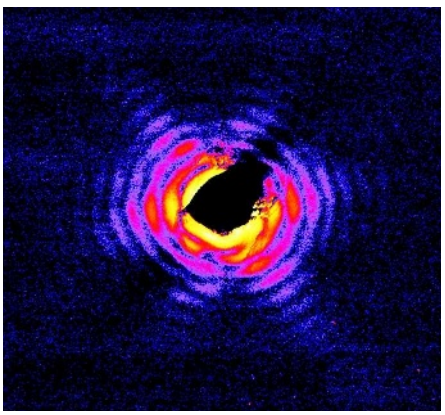
By measuring 14 facial proportions of 312 adults from 52 families, Bereczkei's team shows significant correlations in appearance between young men and their partner's father and young women and their partner's mother. This supports the theory that children are imprinted with their opposite-sex parent's face.

X-RAY PHYSICS

Superman's sharper vision

Phys. Rev. Lett. **101,** 090801 (2008)

X-rays are commonly used to study everything from semiconductors to proteins. But the special optics commonly used to focus these rays struggle to produce images better than a few tens of nanometres in resolution.



C. G. SCHROER *ET AL.*/AM. PHYS. SOC.

Christian Schroer of the Technical University in Dresden, Germany, and his colleagues have improved their X-ray vision by using better beams. A coherent X-ray beam 100 nanometres in diameter produced a diffraction pattern (pictured below left) that could be processed to reveal details of a small gold particle just 5 nanometres across. The technique may be used in future large-scale X-ray facilities.

IMMUNOLOGY

Holistic medicine

PLoS Pathog. **4,** e1000138 (2008)

Immunologists have long known that inactivated whole-virus vaccines are superior to viral-subunit or split-virus vaccines. Anke Huckriede at the University of Groningen in the Netherlands and her colleagues show that for an H5N1 influenza vaccine this enhanced efficacy is due to the action of viral single-stranded RNA molecules. These stimulate the innate immune response, an arm of the immune system that responds quickly and boosts long-term immunity.

The team looked at Toll-like receptors (TLRs), proteins that often initiate innate immune responses. Mice lacking TLR7 — which recognizes the influenza virus's single-stranded RNAs — or other proteins in the same pathway had a degraded immune response to a whole-virus H5N1 vaccine.

ORGANIC CHEMISTRY

Tag-team catalysts

Science doi:10.1126/science.1161976 (2008)

David Nicewicz and David MacMillan at Princeton University in New Jersey have created a double-headed catalytic system to give an aldehyde molecule an alkyl group in a specific position, and with a specific geometry.

Their technique depends on a pincer movement. A ruthenium-based 'photoredox' catalyst that shifts electrons one at a time when hit with fluorescent light

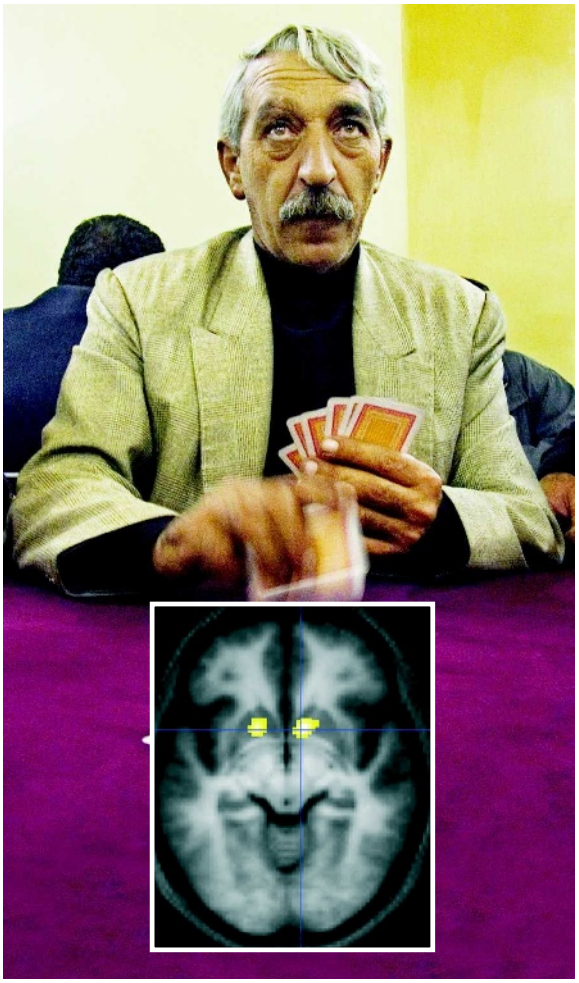
Subliminal choices

Neuron **59,** 561–567 (2008)

Humans can learn to assess risks on the basis of visual hints they are not aware of seeing.

Mathias Pessiglione of the Pitié-Salpêtrière Hospital in Paris and his colleagues repeatedly showed 20 subjects abstract symbols as they played a gambling game. Each symbol presentation involved one of three choices and was followed by a 'masking image' in a series that flickered so fast that the subjects could not consciously perceive the symbol shapes. The subjects were told that the symbols were associated with winning or losing, and then allowed to gamble.

The subjects won more than they lost, indicating that their brains recognized the unperceived symbols and learned to associate them with reward or punishment. Functional neuroimaging showed that the mechanism involves the ventral striatum, a brain area associated with assessing reward value (pictured right).



G. GEORGIOU/PANOS

M. PESSIGLIONE

RESEARCH HIGHLIGHTS

CLIMATE CHANGE

'Hockey stick' holds up

Proc. Natl Acad. Sci. USA **105,** 13252–13257 (2008)

A fresh analysis of climate indicators shows that the Northern Hemisphere is warmer now than it has been in at least 1,300 years.

Previous analyses of climatic history by Michael Mann of Pennsylvania State University in University Park and his colleagues produced a distinctive 'hockey stick' shape; but some of this analysis, and the tree-ring data it used, came under attack.

The latest work by Mann and his co-workers involves various climate proxies, including corals, ice cores, historical records and marine sediments. The authors show that current warming is anomalous even if all tree-ring data are eschewed.

SEXUAL IMPRINTING

Facing Oedipus

Proc. R. Soc. B doi:10.1098/rspb.2008.1021 (2008)

The suggestion that people seek mates that resemble their parents is as old as civilization. Tamas Bereczkei and his colleagues at the University of Pécs in Hungary have found new evidence linking partner choices to parental appearance.

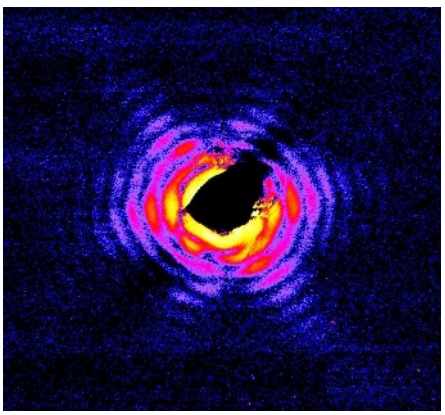
By measuring 14 facial proportions of 312 adults from 52 families, Bereczkei's team shows significant correlations in appearance between young men and their partner's father and young women and their partner's mother. This supports the theory that children are imprinted with their opposite-sex parent's face.

X-RAY PHYSICS

Superman's sharper vision

Phys. Rev. Lett. **101,** 090801 (2008)

X-rays are commonly used to study everything from semiconductors to proteins. But the special optics commonly used to focus these rays struggle to produce images better than a few tens of nanometres in resolution.



C. G. SCHROER *ET AL.*/AM. PHYS. SOC.

Christian Schroer of the Technical University in Dresden, Germany, and his colleagues have improved their X-ray vision by using better beams. A coherent X-ray beam 100 nanometres in diameter produced a diffraction pattern (pictured below left) that could be processed to reveal details of a small gold particle just 5 nanometres across. The technique may be used in future large-scale X-ray facilities.

IMMUNOLOGY

Holistic medicine

PLoS Pathog. **4,** e1000138 (2008)

Immunologists have long known that inactivated whole-virus vaccines are superior to viral-subunit or split-virus vaccines. Anke Huckriede at the University of Groningen in the Netherlands and her colleagues show that for an H5N1 influenza vaccine this enhanced efficacy is due to the action of viral single-stranded RNA molecules. These stimulate the innate immune response, an arm of the immune system that responds quickly and boosts long-term immunity.

The team looked at Toll-like receptors (TLRs), proteins that often initiate innate immune responses. Mice lacking TLR7 — which recognizes the influenza virus's single-stranded RNAs — or other proteins in the same pathway had a degraded immune response to a whole-virus H5N1 vaccine.

ORGANIC CHEMISTRY

Tag-team catalysts

Science doi:10.1126/science.1161976 (2008)

David Nicewicz and David MacMillan at Princeton University in New Jersey have created a double-headed catalytic system to give an aldehyde molecule an alkyl group in a specific position, and with a specific geometry.

Their technique depends on a pincer movement. A ruthenium-based 'photoredox' catalyst that shifts electrons one at a time when hit with fluorescent light

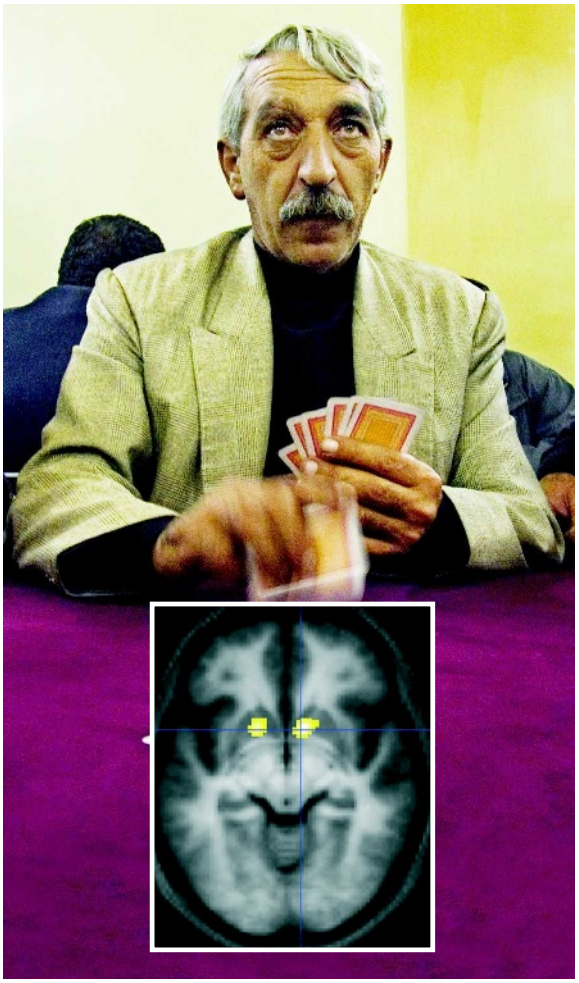
Subliminal choices

Neuron **59,** 561–567 (2008)

Humans can learn to assess risks on the basis of visual hints they are not aware of seeing.

Mathias Pessiglione of the Pitié-Salpêtrière Hospital in Paris and his colleagues repeatedly showed 20 subjects abstract symbols as they played a gambling game. Each symbol presentation involved one of three choices and was followed by a 'masking image' in a series that flickered so fast that the subjects could not consciously perceive the symbol shapes. The subjects were told that the symbols were associated with winning or losing, and then allowed to gamble.

The subjects won more than they lost, indicating that their brains recognized the unperceived symbols and learned to associate them with reward or punishment. Functional neuroimaging showed that the mechanism involves the ventral striatum, a brain area associated with assessing reward value (pictured right).



G. GEORGIOU/PANOS

M. PESSIGLIONE

is one prong; an organocatalyst developed to move single electrons is the other.

The light-activated ruthenium catalyst creates an alkyl halide radical; the aldehyde is dealt with by the organocatalyst; and the reactants are brought together with precision so as to give most of the product the desired handedness. The reaction is easy to perform and broadly applicable, say the authors, and will make life easier for those developing new drugs.

GENETICS

Sweet longevity

Proc. Natl Acad. Sci. USA **105,** 13987–13992 (2008)

Variations in a gene that mediates responses to insulin are associated with longevity in humans, researchers have found.

Bradley Willcox of the Pacific Health Research Institute in Honolulu, Hawaii, and his colleagues looked for links between longevity and variations in five genes involved in insulin signalling and which had previously been suggested to have a link with ageing. The researchers used samples from more than 600 Japanese-American men: 213 who had lived to at least 95 years of age, and 402 who had died before the age of 81.

Variation within one of the genes, *FOXO3A*, was associated with longevity. Those with two copies of a particular version of the gene reported fewer health problems and were nearly three times more likely than those with just one copy to live to the age of 98.

MOLECULAR IMMUNOLOGY

Friendly antibodies

Science 321, 1343–1346 (2008)

The gene *Rfv3* has long been known to protect mice against the 'Friend' retrovirus, but its mechanism has proved elusive.

Warner Greene at the University of California, San Francisco, Kim Hasenkrug at Rocky Mountain Laboratories in Hamilton, Montana, and their co-workers show that the effect comes from *Rfv3*'s role in editing the RNA transcript of the gene *Apobec3*. By intervening in this pathway, they reduced the number of neutralizing antibodies against Friend virus that mice made, and increased their subsequent mortality.

The researchers suggest that this throws light on the importance of the action of the HIV protein Vif on *APOBEC3* in humans, which they think accounts for the poor neutralizing antibody responses generally seen in HIV infection.

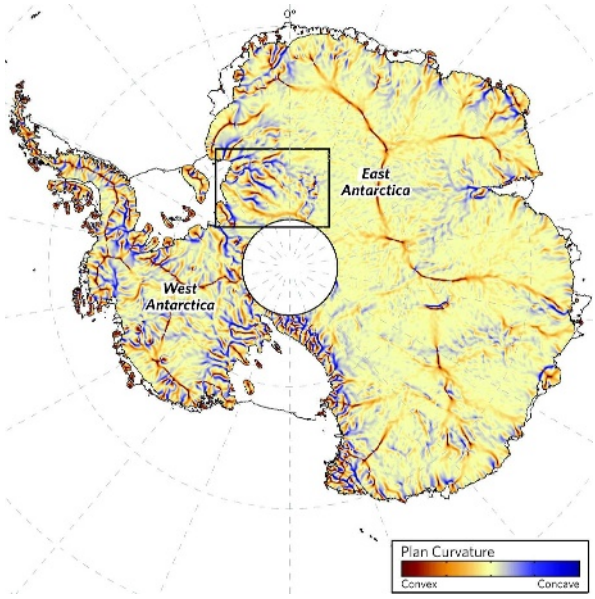
GLACIOLOGY

Judge a basin by its cover

Geophys Res Lett. doi:10.1029/2008GL034728 (2008)

A study of ice-surface shape adds to evidence that a significant part of the East Antarctic Ice Sheet (EAIS) sits on sediments below sea level, which would tend to make it less stable.

Anne Le Brocq of Durham University, UK, and her colleagues studied the 'plan curvature' — a measure of the sinuousness of contours — of the ice sheet. The plan curvature depends on the dynamics of the ice in a way that reveals the depth and deformability of the underlying surface. From this analysis, the researchers suggest that the Recovery Glacier may be sitting on saturated sediments more than 1,000 metres below sea level (area marked on map, below). If this part of the EAIS were lost owing to instability, it would raise sea levels globally by at least 2.6 metres.



A. M. LE BROCQ *ET AL.*/AM. GEOPHYS. UNION

ASTROPHYSICS

Cloudy skies

Astrophys. J. **684,** 364–372 (2008)

Christopher Thom of the University of Chicago in Illinois and his colleagues used a high-resolution spectrometer at the Keck Observatory in Hawaii to look at stars in the direction of a cloud called Complex C to assess its distance from Earth.

The spectra of some stars showed absorption lines that could be ascribed to the cloud, proving that they lay behind it; the spectra of others had no such feature. Because the distances to the stars are known, this provided a bracket for the distance to the cloud. Subsequent estimates of its mass suggest that it could be a significant source of fresh material to the galactic disc.

JOURNAL CLUB

Michael K. Richardson
Leiden University, the Netherlands

A developmental biologist highlights potential pitfalls of using stem cells that can 'remember' their origins.

For me, embryos are beautiful and their development is endlessly fascinating. They are experts at making new tissues, and accomplish this by using stem cells. Stem cells can develop into mature tissues such as bone or muscle; but, cleverly, some of their progeny remain in an undeveloped state, forming reserve supplies that remain in our bodies into adulthood.

Adult stem cells are found in tissues where cell populations are constantly being renewed, such as the testes, hair follicles and bones. We replace our entire skeleton every decade or so, and rely on stem cells in our bones to do this. Stem cells also have an important role in repair, swinging into action to deal with broken bones and other mishaps.

A recent study in mice yielded remarkable evidence that some of these adult stem cells remember where in the embryo they came from. Jill Helms and her colleagues at Stanford University in California grafted stem cells from one bone into another to see whether they would help repair fractures in the 'wrong' location. Stem cells transplanted from leg bones into fractured jaws failed to produce new bone (P. Leucht *et al. Development* **135,** 2845–2854; 2008). Interestingly, the uncooperative stem cells continued to express a gene, *Hoxa11*, that acts as a kind of embryonic 'postcode' for the leg.

These findings have broad implications. They validate the concept of non-equivalence — that seemingly identical cells differ if they come from different places in the embryo — first enunciated by Julian Lewis and Lewis Wolpert in the 1970s, and show that it holds in the adult. More pragmatically, if some stem cells also have positional memory, doctors may need to make sure that they take stem cells from the right location to heal damaged tissues.

Discuss this paper at **http://blogs.nature.com/nature/journalclub**

is one prong; an organocatalyst developed to move single electrons is the other.

The light-activated ruthenium catalyst creates an alkyl halide radical; the aldehyde is dealt with by the organocatalyst; and the reactants are brought together with precision so as to give most of the product the desired handedness. The reaction is easy to perform and broadly applicable, say the authors, and will make life easier for those developing new drugs.

GENETICS

Sweet longevity

Proc. Natl Acad. Sci. USA **105,** 13987–13992 (2008)

Variations in a gene that mediates responses to insulin are associated with longevity in humans, researchers have found.

Bradley Willcox of the Pacific Health Research Institute in Honolulu, Hawaii, and his colleagues looked for links between longevity and variations in five genes involved in insulin signalling and which had previously been suggested to have a link with ageing. The researchers used samples from more than 600 Japanese-American men: 213 who had lived to at least 95 years of age, and 402 who had died before the age of 81.

Variation within one of the genes, *FOXO3A*, was associated with longevity. Those with two copies of a particular version of the gene reported fewer health problems and were nearly three times more likely than those with just one copy to live to the age of 98.

MOLECULAR IMMUNOLOGY

Friendly antibodies

Science 321, 1343–1346 (2008)

The gene *Rfv3* has long been known to protect mice against the 'Friend' retrovirus, but its mechanism has proved elusive.

Warner Greene at the University of California, San Francisco, Kim Hasenkrug at Rocky Mountain Laboratories in Hamilton, Montana, and their co-workers show that the effect comes from *Rfv3*'s role in editing the RNA transcript of the gene *Apobec3*. By intervening in this pathway, they reduced the number of neutralizing antibodies against Friend virus that mice made, and increased their subsequent mortality.

The researchers suggest that this throws light on the importance of the action of the HIV protein Vif on *APOBEC3* in humans, which they think accounts for the poor neutralizing antibody responses generally seen in HIV infection.

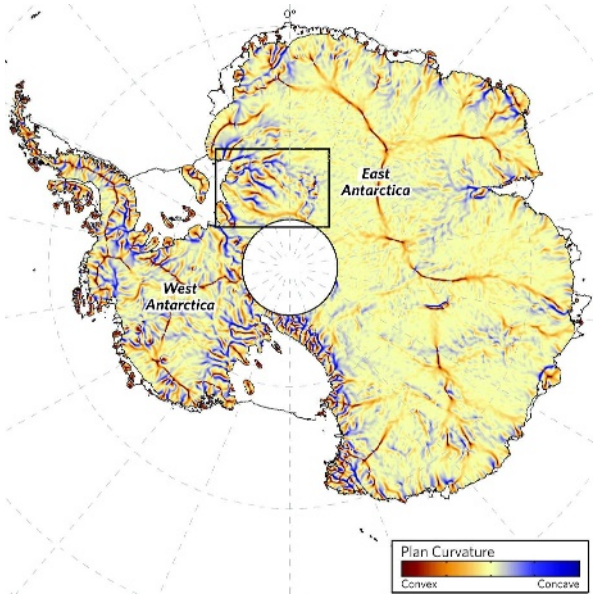
GLACIOLOGY

Judge a basin by its cover

Geophys Res Lett. doi:10.1029/2008GL034728 (2008)

A study of ice-surface shape adds to evidence that a significant part of the East Antarctic Ice Sheet (EAIS) sits on sediments below sea level, which would tend to make it less stable.

Anne Le Brocq of Durham University, UK, and her colleagues studied the 'plan curvature' — a measure of the sinuousness of contours — of the ice sheet. The plan curvature depends on the dynamics of the ice in a way that reveals the depth and deformability of the underlying surface. From this analysis, the researchers suggest that the Recovery Glacier may be sitting on saturated sediments more than 1,000 metres below sea level (area marked on map, below). If this part of the EAIS were lost owing to instability, it would raise sea levels globally by at least 2.6 metres.



A. M. LE BROCQ *ET AL.*/AM. GEOPHYS. UNION

ASTROPHYSICS

Cloudy skies

Astrophys. J. **684,** 364–372 (2008)

Christopher Thom of the University of Chicago in Illinois and his colleagues used a high-resolution spectrometer at the Keck Observatory in Hawaii to look at stars in the direction of a cloud called Complex C to assess its distance from Earth.

The spectra of some stars showed absorption lines that could be ascribed to the cloud, proving that they lay behind it; the spectra of others had no such feature. Because the distances to the stars are known, this provided a bracket for the distance to the cloud. Subsequent estimates of its mass suggest that it could be a significant source of fresh material to the galactic disc.

JOURNAL CLUB

Michael K. Richardson
Leiden University, the Netherlands

A developmental biologist highlights potential pitfalls of using stem cells that can 'remember' their origins.

For me, embryos are beautiful and their development is endlessly fascinating. They are experts at making new tissues, and accomplish this by using stem cells. Stem cells can develop into mature tissues such as bone or muscle; but, cleverly, some of their progeny remain in an undeveloped state, forming reserve supplies that remain in our bodies into adulthood.

Adult stem cells are found in tissues where cell populations are constantly being renewed, such as the testes, hair follicles and bones. We replace our entire skeleton every decade or so, and rely on stem cells in our bones to do this. Stem cells also have an important role in repair, swinging into action to deal with broken bones and other mishaps.

A recent study in mice yielded remarkable evidence that some of these adult stem cells remember where in the embryo they came from. Jill Helms and her colleagues at Stanford University in California grafted stem cells from one bone into another to see whether they would help repair fractures in the 'wrong' location. Stem cells transplanted from leg bones into fractured jaws failed to produce new bone (P. Leucht *et al. Development* **135,** 2845–2854; 2008). Interestingly, the uncooperative stem cells continued to express a gene, *Hoxa11*, that acts as a kind of embryonic 'postcode' for the leg.

These findings have broad implications. They validate the concept of non-equivalence — that seemingly identical cells differ if they come from different places in the embryo — first enunciated by Julian Lewis and Lewis Wolpert in the 1970s, and show that it holds in the adult. More pragmatically, if some stem cells also have positional memory, doctors may need to make sure that they take stem cells from the right location to heal damaged tissues.

Discuss this paper at **http://blogs.nature.com/nature/journalclub**

NEWS

China rushes through major funding system

Next month, China will start pouring money into two long-awaited 'megaprojects' in infectious disease and drug discovery. But although scientists welcome the funding bonanza, many criticize how it is being administered — especially because, after waiting more than two years for the announcement, scientists were given only a few weeks to apply.

In February 2006, the national medium- and long-term programme for science and technology development, which lays out plans up to 2020, called for investments in biomedical sciences. After debate among the ministries of health, science and technology, and industry — along with the powerful National Development and Reform Commission — it was decided that the health ministry would take the lead. But the negotiations took so long that requests for grant applications did not go out until the middle of last month.

Funding levels are not yet set, but insiders say roughly 6 billion yuan (US$880 million) will be allocated for drug discovery and 3 billion yuan for infectious disease research. "The funding is huge — an unprecedented amount," says Ray Yip, who works on infectious disease at the Bill & Melinda Gates Foundation in Beijing. In addition to university and institute researchers, industrial researchers can apply if their company is majority Chinese-owned.

Projects that win funding will run between October 2008 and December 2010. It is still unclear, however, whether all of the funding earmarked for this first stage will be squeezed into the next two years or allowed to carry over to the next Five Year Plan.

The infectious-disease money will go to research on HIV/AIDS, hepatitis B, hepatitis C and tuberculosis. The drug-discovery money will target ten major diseases including cancer, cardiovascular disease, neurodegenerative diseases, diabetes and mental illness. The biggest chunks of cash, however, will go towards the establishment of Good Laboratory Practice platforms — in which scientists establish procedures to ensure the accuracy of safety data of laboratory materials — and Good Clinical Practice platforms, which ensure the rights and the safety of patients involved in clinical trials. "The number-one goal is to make similar regulations to those in the United States and elsewhere," says Wei He, an immunologist at the Chinese Academy of Medical Sciences in Beijing, who was pulled in to organize the review of the applications in August and hasn't had a day off since.

Yip says that the financial lure of the megaprojects might help remedy China's shortage of experienced biomedical researchers. "It will no doubt attract talented people," he says. "Those who didn't want to go back because the grants were too small will now find that they can get several million dollars to do a research project. The tables will start turning."

The projects are being rushed through in a post-Olympics scurry. The deadline for infectious disease applications was 31 August, just 16 days after the request for applications went up on the health ministry website. Likewise, researchers had only three weeks to meet the

Reaction to China's new biomedical funding programmes has been mixed.

GUANG NIU/GETTY IMAGES

Megacity project seeks to gauge urban pollution

Parisian pollution is nothing compared with that of Beijing or Mexico City. Yet Paris, with 11 million people crammed into a region just 20 kilometres across, is to take centre stage in a new research project on the impact of megacities on air pollution.

The MEGAPOLI project, which starts next month, will focus on building regional air-pollution models for every city in the world with a population of more than five million. It encompasses 23 research organizations from 11 European countries, along with 24 collaborating partners outside Europe. Project scientists hope that it will lead to better maps of potential exposure to harmful aerosols and particles, and to improved urban planning.

Two dozen cities worldwide already have populations exceeding 10 million; much of the existing field data on what these do to the air come from studies of individual cities or regions. The novelty of MEGAPOLI is to try to understand the bigger picture of the impact of megacities and the feedbacks between pollution and climate, says Alexander Baklanov of the Danish Meteorological Institute, a project leader. That includes, he says, studying atmospheric processes "from the street scale and up to the global scale, with interactions in both directions".

This colour of this balloon indicates Paris's air quality.

AÉROPHILE

For all its grand goals, MEGAPOLI is budgeted at a relatively cheap €3 million (US$4.3 million). That is because the project is 80% modelling and 20% measurements, says Mark Lawrence, a chemical transport modeller at the Max Planck Institute for Chemistry in Mainz, Germany. Much of the data have already been collected by its collaborating partners; MEGAPOLI will work, for instance, with the Milagro project, which carried out the largest air-pollution field campaign to date, in Mexico City in 2006.

MEGAPOLI will work with similar data sets from 11 other megacities,

11 September deadline for applications to the drug-discovery project. The review process will be similarly speedy, taking between ten days and a fortnight. Distribution of funds will start in October.

Many scientists are angry at being hurried, although they shy away from publicly criticizing the project while their applications are under review. One molecular biologist in Guangzhou called the last-minute rush "ridiculous", saying it would reduce the quality of proposals. Noting that some proposals had already been rejected without any explanation, he said the review is "like a black hole".

Although the application forms claim the procedure will include "public announcement, free application, expert review [and] merit-based selection", some wonder whether the speedy application and evaluation procedures mean the winners have already been picked. "It only benefits the people who knew about it long before everyone else," says the head of a Shanghai biotech company. Even those not critical of the project say the money will go to the usual suspects. Others say the ministry should have taken more time explaining the projects and their goals to those not in the inner circle.

A senior biologist in Beijing criticizes the focus on hepatitis B, given that a vaccine already exists. He says that he wonders whether the money might be more effective if it were split between vaccination programmes and other research programmes. "These megaprojects are covers for dividing up funds, not driven by real goals," he says.

There is also concern about whether the funding will be spread too thinly. "It will help everyone a little bit, but not have a big impact on new drug development," says a researcher at the Shanghai Institute of Materia Medica. Yip is more positive about the infectious-disease money. "Even if they spread it around," he says, "there is still a substantial amount."

Some scientists contacted by *Nature* said they could easily repackage their existing research for the megaprojects. Others brush aside criticisms, noting that those familiar with the Chinese funding system should have been ready. Results of the selection process are expected later this month. ■

David Cyranoski

including Beijing, Mumbai, New York and Tokyo. "We are not starting from a blank page," says Baklanov. The comprehensive datasets will be used to build regional models, which will be interfaced with less detailed data — mainly global-scale models and satellite data of both air pollution and climate.

To complete the picture, the consortium will model four European metropolitan areas: Paris, London, Germany's Rhine-Ruhr region and the Po valley in Italy. Paris will be studied in the most detail, with an aircraft and ground field campaign to plug gaps in existing air-pollution data — particularly in the chemical speciation and evolution of aerosols, as well as gas-aerosol interactions. It will also benefit from the results of a second EU-led megacity project, CityZen, which will focus on determining the distribution and changes in air pollution over the past decade in four hotspots. The result, says Baklanov, will not only refine models and maps, but also tools to help urban planners mitigate pollution.

Studying many megacities together is crucial to building better regional models, says Jeffrey Gaffney, an atmospheric chemist at the University of Arkansas in Little Rock, who is not involved in the project. Moreover, he says, as cities worldwide differ in how they deal with pollution, studying many cities will itself provide benchmarks and better predictions of what works best in improving urban management of emissions.

"The collaborative approach in MEGAPOLI is a good one," he notes. "By combining efforts, the sum of the instrumentation, expertise and quality of the data is greater than any one investigator could ever hope to mount." ■

Declan Butler

See Editorial, page 137.

LAB POLITICS

GARRISON PHOTOGRAPHY

In the second of our election-themed podcasts available online, *Nature* looks at where US biomedical research might head after November's presidential election. Excerpts from our panel discussion:

"How are we going to structure our biomedical research enterprise, our graduate training and our undergraduate training for the next generation of scientists? Republicans and Democrats should be able to pull in the same direction on these issues."
Thomas Cech, president, Howard Hughes Medical Institute, Chevy Chase, Maryland

"The prohibition on federal funding of most human embryonic stem-cell research has been an enormous wet blanket on the whole research enterprise in this area."
Jonathan Moreno, University of Pennsylvania, Philadelphia

"[Stem-cell research] has become so politicized, and that has encouraged some scientists to become very exuberant about the potential. Whereas if it hadn't become so politicized, I think they would be a bit more sceptical."
Thomas Cech

"We must preserve the synergy that we have between the public and the private sectors, if we intend to maintain our competitive lead in science and technology."
Gail Cassell, vice-president for scientific affairs, Eli Lilly, Indianapolis, Indiana

"It might even be time for there to be a life scientist as the science adviser to the president, which would be a departure."
Jonathan Moreno

To hear the full discussion, chaired by our columnist David Goldston, visit www.nature.com/nature/podcast. Next week's instalment: innovation policy.

11 September deadline for applications to the drug-discovery project. The review process will be similarly speedy, taking between ten days and a fortnight. Distribution of funds will start in October.

Many scientists are angry at being hurried, although they shy away from publicly criticizing the project while their applications are under review. One molecular biologist in Guangzhou called the last-minute rush "ridiculous", saying it would reduce the quality of proposals. Noting that some proposals had already been rejected without any explanation, he said the review is "like a black hole".

Although the application forms claim the procedure will include "public announcement, free application, expert review [and] merit-based selection", some wonder whether the speedy application and evaluation procedures mean the winners have already been picked. "It only benefits the people who knew about it long before everyone else," says the head of a Shanghai biotech company. Even those not critical of the project say the money will go to the usual suspects. Others say the ministry should have taken more time explaining the projects and their goals to those not in the inner circle.

A senior biologist in Beijing criticizes the focus on hepatitis B, given that a vaccine already exists. He says that he wonders whether the money might be more effective if it were split between vaccination programmes and other research programmes. "These megaprojects are covers for dividing up funds, not driven by real goals," he says.

There is also concern about whether the funding will be spread too thinly. "It will help everyone a little bit, but not have a big impact on new drug development," says a researcher at the Shanghai Institute of Materia Medica. Yip is more positive about the infectious-disease money. "Even if they spread it around," he says, "there is still a substantial amount."

Some scientists contacted by *Nature* said they could easily repackage their existing research for the megaprojects. Others brush aside criticisms, noting that those familiar with the Chinese funding system should have been ready. Results of the selection process are expected later this month. ■

David Cyranoski

including Beijing, Mumbai, New York and Tokyo. "We are not starting from a blank page," says Baklanov. The comprehensive datasets will be used to build regional models, which will be interfaced with less detailed data — mainly global-scale models and satellite data of both air pollution and climate.

To complete the picture, the consortium will model four European metropolitan areas: Paris, London, Germany's Rhine-Ruhr region and the Po valley in Italy. Paris will be studied in the most detail, with an aircraft and ground field campaign to plug gaps in existing air-pollution data — particularly in the chemical speciation and evolution of aerosols, as well as gas-aerosol interactions. It will also benefit from the results of a second EU-led megacity project, CityZen, which will focus on determining the distribution and changes in air pollution over the past decade in four hotspots. The result, says Baklanov, will not only refine models and maps, but also tools to help urban planners mitigate pollution.

Studying many megacities together is crucial to building better regional models, says Jeffrey Gaffney, an atmospheric chemist at the University of Arkansas in Little Rock, who is not involved in the project. Moreover, he says, as cities worldwide differ in how they deal with pollution, studying many cities will itself provide benchmarks and better predictions of what works best in improving urban management of emissions.

"The collaborative approach in MEGAPOLI is a good one," he notes. "By combining efforts, the sum of the instrumentation, expertise and quality of the data is greater than any one investigator could ever hope to mount." ■

Declan Butler

See Editorial, page 137.

LAB POLITICS

GARRISON PHOTOGRAPHY

In the second of our election-themed podcasts available online, *Nature* looks at where US biomedical research might head after November's presidential election. Excerpts from our panel discussion:

"How are we going to structure our biomedical research enterprise, our graduate training and our undergraduate training for the next generation of scientists? Republicans and Democrats should be able to pull in the same direction on these issues."
Thomas Cech, president, Howard Hughes Medical Institute, Chevy Chase, Maryland

"The prohibition on federal funding of most human embryonic stem-cell research has been an enormous wet blanket on the whole research enterprise in this area."
Jonathan Moreno, University of Pennsylvania, Philadelphia

"[Stem-cell research] has become so politicized, and that has encouraged some scientists to become very exuberant about the potential. Whereas if it hadn't become so politicized, I think they would be a bit more sceptical."
Thomas Cech

"We must preserve the synergy that we have between the public and the private sectors, if we intend to maintain our competitive lead in science and technology."
Gail Cassell, vice-president for scientific affairs, Eli Lilly, Indianapolis, Indiana

"It might even be time for there to be a life scientist as the science adviser to the president, which would be a departure."
Jonathan Moreno

To hear the full discussion, chaired by our columnist David Goldston, visit www.nature.com/nature/podcast. Next week's instalment: innovation policy.

Change at the top for climate panel

The Intergovernmental Panel on Climate Change (IPCC) elected new leadership last week in Geneva, Switzerland, unanimously re-electing the Indian economist Rajendra Pachauri as chairman but choosing new heads of the three working groups that coordinate the writing of its massive reports.

The most significant race was for the leadership of Working Group I, which evaluates the basic science of climate change. Swiss climate modeller Thomas Stocker came out ahead after two initial votes narrowed a field of four to him and Francis Zwiers, a senior scientist at the Canadian Centre for Climate Modelling and Analysis in Toronto.

Stocker heads the climate and environmental physics unit at the University of Bern and has been a coordinating lead author in Working Group I during the past two assessments. He sees plenty of work to do in assessing sea-level rise, carbon-cycle feedbacks, and regional impacts. Qin Dahe, who heads the China Meteorological Association, stays on as co-chair representing a developing nation.

"The value of the IPCC is utterly dependent on top scientists such as these as co-chairs to lead the assessment process," says Susan Solomon of the US National Oceanic and Atmospheric Administration in Boulder, Colorado, who served as lead co-chair of Working Group I for the most recent assessment in 2007.

Working Group II, which assesses impacts and adaptation, will be chaired by Chris Field, an ecologist who heads the Carnegie Institution's Department of Global Ecology in Stanford, California. His co-chair will be Vicente Barros, an oceanographer at the University of Buenos Aires.

J. NEUHAUS

Newly elected co-chairs of IPCC working groups Ottmar Edenhofer (left) and Chris Field.

For the next assessment, Field says he wants to further integrate the sciences pertaining to impacts and adaptation. He also wants to drill down to "the levels of processes" instead of just listing the types of changes that can be expected under climate change. He will also push for a separate chapter on oceans to look at acidification, warming and loss of ice cover.

Working Group III, which assesses mitigation options, will be headed by Ottmar Edenhofer, chief economist at the Potsdam Institute for Climate Impact Research in Germany. He says he wants to give business people and policy-makers a larger role and "be a bit more precise" about the advantages, costs and risks of different options.

All the co-chairs support further integration among the three working groups. They differ in their attitude to special reports — targeted assessments to address pressing policy questions. Edenhofer favours their use; Field and Stocker don't emphasize them as much. The panel is currently preparing a special report on renewable energy, and Edenhofer says he has secured an agreement from the German government for additional resources for such purposes.

In a departure forced by a tight vote and procedural complications, Edenhofer will have two co-chairs: Ramón Pichs Madruga, an economist at the Center for Research on the World Economy in Havana, and Youba Sokona, a Malian environmental scientist who heads the Sahara and Sahel Observatory in Tunis.

"It was a political compromise allowing everybody to save face," says new IPCC vice-chair Jean-Pascal Van Ypersele, a climatologist at the Catholic University of Louvain, Belgium. "It was not perfect, but it in the end it was accepted by everybody." ■

Jeff Tollefson

Spending the Nobel prize

Since winning the Nobel Peace Prize in 2007, the Intergovernmental Panel on Climate Change (IPCC) has had 885,000 Swiss francs (US$785,000) burning a hole in its pocket.

Now it proposes to put the money into a trust fund, tentatively named the IPCC Bert Bolin Memorial Scholarship Fund, after the Swedish climatologist who was the panel's first chairman and who died on 30 December last year. The fund would support PhD students and postdocs from the developing world.

Organizers expect private donors to supplement the prize money. The IPCC chairman and working-group co-chairs will pick scholars, with the aim of building understanding and management of climate change in developing nations, where its impacts are expected to be highest.

Saleemul Huq, head of the International Institute for Environment and Development's climate-change programme in London, believes the fund will also benefit the IPCC. "There's still a dearth of qualified scientists in developing countries who can be drawn on as lead authors," he says.

But scholarships are no radical innovation, says Mickey Glantz, director of the Center for Capacity Building, currently at the US National Center for Atmospheric Research in Boulder, Colorado.

"The developing world has a lot of expertise in it, but we keep them in a dependent mode," he says. "We should be fostering south–south connections, not just north–south connections." **Anna Barnett**

French university under fire for culling macaques

Primate scientists are criticizing a decision at the Louis Pasteur University in Strasbourg, France, to kill a research colony of Tonkean macaques (*Macaca tonkeana*) last month because the animals were infected with the herpes B virus.

The monkeys, at the Centre of Primatology, had never shown symptoms of disease, and scientists critical of the move say that the culling was scientifically and morally unjustified. But university officials say they were concerned that the virus could jump the species barrier to people working with the animals. In humans, the virus can cause fatal encephalomyelitis.

Some of the animals in the 14-strong colony had been used for up to 25 years by macaque ethologists. The species has an unusual way of resolving social conflicts, in which a third individual often tries to actively reconcile two fighting monkeys. There are few Tonkean macaques in captivity.

It has been known since the 1980s that the Strasbourg animals carried the virus, but it was not until more recently that the possibility of human contagion was realized. Safety protocols were implemented in 1998, when the colony was isolated in its own wooded enclosure. Scientists and keepers were required to wear protective clothing when entering the enclosure, although they rarely get close to the animals. Restrictions became tighter when a new veterinary surgeon joined the university in 2002 and research students were not allowed to enter. A new colony of 23 animals was bred from virus-free individuals.

The university has asked its primate ethologists, led by Bernard Thierry, not to speak to the press, but their colleagues elsewhere say that they were distressed to learn that the animals were killed without their knowledge on Sunday 31 August, when scientists and keepers were not present. "We were all shocked by this," says Elisabetta Visalberghi, a researcher at the Institute of Cognitive Sciences and Technologies in Rome, and president of the Italian Ethological Society.

The ethologists had been negotiating with the university for more than five years over the animals' fate, including talks with a sanctuary in San Antonio, Texas, over possibly accepting the macaques. But on 29 August, a university council confirmed a decision taken earlier in the summer to cull the animals.

G. CUBITT/NHPA/PHOTOSHOT

Herpes B is common in Tonkean macaques.

"It would be stupid to keep them for more time," says Nicolas Herrenschmidt, director of the centre. "The risk of transmission to humans is small, but it is there."

Juichi Yamagiwa at Kyoto University in Japan, who is president of the International Primatological Society, says macaque-to-human transmission of the virus is very rare, and the society has guidelines to ensure this does not happen. "I believe that no one would insist on killing infected monkeys if they have read these guidelines," he says.

Behaviouralist Frans de Waal works with macaques, many of which are infected, at Emory University's Yerkes National Primate Research Center in Atlanta, Georgia. He says he is "shocked that the deed has been done". He believes that "the risk, if managed properly, is not great enough to justify euthanizing these beautiful and interesting animals".

Hannah Buchanan-Smith, a primate behaviouralist at the University of Stirling, UK, says that the cull was "morally unacceptable". Research primates are not expendable once they have stopped being useful, she says. "We have a moral responsibility to look after them once we have used them, if, like these, they were able to lead happy lives."

But Alain Beretz, president of the Louis Pasteur University, says that the decision to kill the animals was made in a considered fashion over a five-year period. "We were not happy with the decision that we had to make," he says. "but we had to do it to protect our employees." ■

Alison Abbott

CERN

LARGE HADRON COLLIDER
World's biggest particle accelerator fires up.
www.nature.com/news

'Lucky' Louisiana unprepared for Gustav

Hurricane Gustav, which made landfall just west of New Orleans on 1 September, had far less devastating effects than Hurricane Katrina three years earlier — on either the people or the land on which they live. But the third major hurricane to hit Louisiana's fragile wetlands in three years has made it clear that, although coastal recovery is high on the state's agenda, little has been done on the ground since 2005.

Just last month, Louisiana governor Bobby Jindal announced that the state would chip in $300 million of its surplus funds for coastal restoration and flood protection. It is "the largest single commitment to coastal restoration ever made by any governor in Louisiana", says Chris Macaluso, a spokesman from the Governor's Office of Coastal Activities.

The funding swelled a pot of post-Katrina restoration money that had been secured, but mostly not spent, in time for Gustav. "Money started coming in, and what happened was another freakin' hurricane hit the coast," says Mark Kulp, a coastal geologist at the University of New Orleans in Louisiana. "Three years is not a lot of time to implement with all the bureaucracy."

In fact, one of the reasons Gustav did less ecological damage than expected may have been that it passed over the state's single remaining large barrier island, says Robert Twilley, a wetland ecologist at Louisiana State University in Baton Rouge. The intact Grand Isle acted as a buffer. "They were lucky," he says.

Still, Gustav has shown how the hurricane cycle is substantially shorter than the bureaucratic cycle of projects that are meant to make the coast healthier and less vulnerable to hurricanes.

Louisiana's coast is riddled with navigation and oil and gas pipeline channels that nibble away at the land from within. Levees and other flood-protection schemes along the Mississippi mean that the region no longer receives regular deposits of sediment from the river. An estimated 62 square kilometres of wetlands vanish each year simply from erosion.

Storms take even more: Katrina and Rita, which both hit in 2005, wiped out 560 square kilometres. But this also convinced many officials that restoring such land is important for protecting settlements, including New Orleans, from the naked fury of a hurricane storm surge.

Before Katrina and Rita, the state and federal government together spent about $75 million a year on coastal restoration. A congressional act later that year secured an additional $510 million over four years for both restoration and hurricane protection. In 2007, then-state governor Kathleen Blanco assigned $200 million of a state surplus to coastal restoration and hurricane protection.

> "Money started coming in, and what happened was another freakin' hurricane hit the coast."

But the money flows slowly through a gummy bureaucracy. In the case of the new Jindal money, Macaluso says, the governor couldn't commit any spare funds to coastal restoration until he had renegotiated with the federal government on repaying a large flood-protection grant.

USGS

Erosion and hurricane damage have taken their toll on the Louisiana barrier island of East Timbalier.

Although hurricane protection and coastal restoration are increasingly being seen as two sides of the same coin, people always trump land. "Because of the demand to fix the weak links in flood protection, a lot of the money spent post-Katrina — almost all of the money that has been made concrete — has gone into flood protection," notes Donald Boesch, president of the University of Maryland Center for Environmental Science in Cambridge and an expert on his native Louisiana coast. That makes sense, he says, but "it is going to be necessary not to lose sight of coastal restoration, or you'll end up with fortresses sitting in the Gulf of Mexico with no landscape around them".

Some restoration projects that predate Katrina include freshwater diversions from the Mississippi and pumping projects to spread sediment onto areas that are eroding. More ambitious, post-Katrina projects are mostly still on paper, including a substantial diversion–pipeline combination project near the hamlet of Myrtle Grove, up the Mississippi River from New Orleans.

There are some positive signs. The state finally has a comprehensive "master plan for a sustainable coast", adopted in June 2007. And a traditional over-reliance on hard structures such as levees, often introduced at the expense of wetlands, may be waning. In June, the US Army Corps of Engineers deauthorized the navigational channel known as MRGO (the Mississippi River Gulf Outlet); the structure is set to be closed off with 355,000 tonnes of rock. And the massive planned Morganza-to-the-Gulf levee, approved last year by Congress, is being re-evaluated after cost estimates ballooned to more than $11 billion.

Still, Kerry St. Pé, director of the Barataria–Terrebonne National Estuary Program in Thibodaux, has been agitating for eight years for a pipeline that would pump sediment from the Atchafalaya River, which flows out of the Mississippi, and from offshore, to rebuild wetlands in his area. It hasn't materialized, even after Katrina, Rita and now Gustav, which badly damaged St. Pé's office roof. There is impatience in his voice when he says: "The one thing we know how to do in Louisiana is lay pipeline."

St. Pé may have to wait through another hurricane first. As *Nature* went to press, Hurricane Ike was looking likely to head into the Gulf of Mexico and strike somewhere in Texas or Louisiana as a major storm. ■

Emma Marris

Cancer complexity slows quest for cure

Hopes that large studies of cancer genomics will justify their high cost by offering a fast-track to cures have been dealt a blow by a series of papers.

The controversial Cancer Genome Atlas, run by the US National Institutes of Health, is analysing genetic and epigenetic changes in cancers. Still in its pilot phase, the project could eventually cost up to $1.35 billion. But it does not include 'functional' studies that investigate how the mutations aid tumour development and what drugs might target the pathways essential to tumour survival.

Yet functional studies are exactly what two papers, from a group led by Bert Vogelstein of the Johns Hopkins School of Medicine in Baltimore, Maryland, suggest are needed. He and others say that the focus should shift from hunting for individual genes that cause certain cancers, to disrupting the broader biological pathways that support cancer growth.

"It is apparent from studies like ours that it is going to be even more difficult than expected to derive real cures," says Vogelstein.

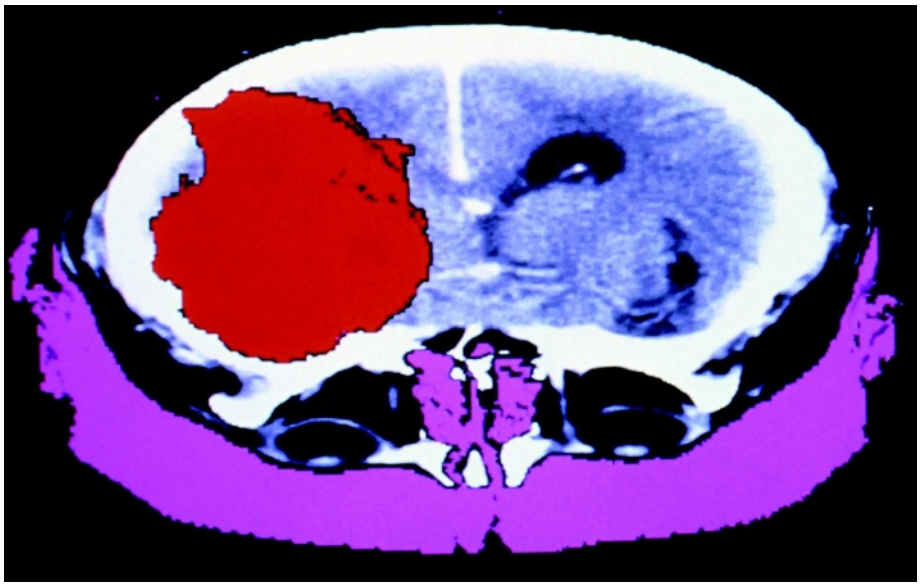
Two papers from his team, plus one from the Cancer Genome Atlas group, appeared on 4 September in *Science*[1,2] and *Nature*[3]. The studies find that individual cancer patients each carry dozens of genetic mutations — an average of 63 alterations in pancreatic cancer and 47 DNA mutations in one type of brain cancer. Similar results have been found in previous studies of other cancers. This makes it unlikely that the cancers will be cured by drugs that target just one or a few genes, the researchers say.

Other scientists call the latest results sobering but important. "Unfortunately this wasn't what we all hoped for," says Stephen Elledge of Brigham and Women's Hospital in Boston, Massachusetts. "But there is useful information in there, and what they're learning is more of what they already learned — which is that cancers are extremely complex."

"What they're learning is more of what they already learned — which is that cancers are extremely complex."

Vogelstein's team analysed nearly 21,000 genes in some 20 patients with a type of brain cancer called glioblastoma multiforme[1] and a similar number with pancreatic cancer[2]. The Cancer Genome Atlas team looked at 600 genes in 91 patients with glioblastoma multiforme[3].

The studies hoped to find genetic glitches driving the tumours that could be treated by drugs similar to Novartis's Gleevec (imatinib) and Genentech's Tarceva (erlotinib), which inhibit the activity of some mutated genes that can cause certain cancers. Instead, the findings confirm earlier hints that patients with the same cancer diagnosis can harbour different sets of genetic causes.



CNRI/SPL

Solid tumours such as glioblastoma (red) can be caused by multiple genes in different patients.

"It is extremely unlikely that drugs that target a single gene, such as Gleevec, will be active against a major fraction of solid tumours," says Vogelstein, whose group has published similar studies on breast and colorectal cancers[4].

The latest papers do identify some single genes that seem to be important in subsets of the cancers. For instance, Vogelstein and his colleagues report that a gene called *IDH1*, which had not previously been linked to brain cancer, was often mutated in younger patients with a certain type of glioblastoma. And the Cancer Genome Atlas group reports that a gene called *NF1*, whose link to cancer had been hypothesized before, was mutated in 23% of 206 patients that were analysed.

The atlas group also found that patients with a particular epigenetic make-up who are treated with one type of chemotherapy show a pattern of genetic and epigenetic changes that may render them resistant to further treatments. The mutation pattern suggests why this resistance evolves and may help doctors find strategies to avoid it.

That shows the importance of the atlas and other similar studies, such as those included in the 10-nation International Cancer Genome Consortium, says Lynda Chin of the Dana-Farber Cancer Institute in Boston, who was the team leader for the atlas paper. "These are very important, clinically relevant questions, and you can't answer them in a traditional hypothesis-driven manner," Chin says.

The Cancer Genome Atlas has moved more slowly than its architects had hoped because it has proved harder than expected to find enough high-quality tissue samples to analyse. The *Nature* paper is an interim analysis of the samples it has studied so far. But most of the genes identified in it and the Vogelstein studies had already been identified in earlier studies, and are not found in most patients with a particular tumour.

That bolsters critics of the atlas, such as Elledge, who has long said that functional studies will be needed to filter the most clinically relevant drug targets out of the massive pool of mutations found in cancer.

"The information they're getting is useful," he says, "but it's expensive and I think some of that money should go to help get you further along into finding drugs." ■

Erika Check Hayden

1. Parsons, D. W. *et al. Science* doi:10.1126/science.1164382 (2008).
2. Jones, S. *et al. Science* doi:10.1126/science.1164368 (2008).
3. The Cancer Genome Atlas Research Network *Nature* doi:10.1038/nature07385 (2008).
4. Sjöblom, T. *et al. Science* **314,** 268–274 (2006).

See Editorial, page 138.

Cancer complexity slows quest for cure

Hopes that large studies of cancer genomics will justify their high cost by offering a fast-track to cures have been dealt a blow by a series of papers.

The controversial Cancer Genome Atlas, run by the US National Institutes of Health, is analysing genetic and epigenetic changes in cancers. Still in its pilot phase, the project could eventually cost up to $1.35 billion. But it does not include 'functional' studies that investigate how the mutations aid tumour development and what drugs might target the pathways essential to tumour survival.

Yet functional studies are exactly what two papers, from a group led by Bert Vogelstein of the Johns Hopkins School of Medicine in Baltimore, Maryland, suggest are needed. He and others say that the focus should shift from hunting for individual genes that cause certain cancers, to disrupting the broader biological pathways that support cancer growth.

"It is apparent from studies like ours that it is going to be even more difficult than expected to derive real cures," says Vogelstein.

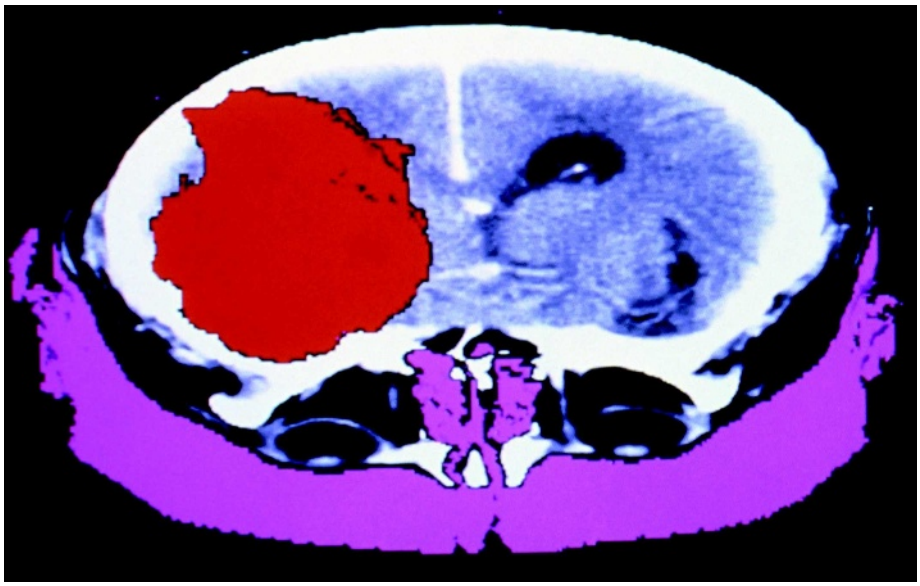
Two papers from his team, plus one from the Cancer Genome Atlas group, appeared on 4 September in *Science*[1,2] and *Nature*[3]. The studies find that individual cancer patients each carry dozens of genetic mutations — an average of 63 alterations in pancreatic cancer and 47 DNA mutations in one type of brain cancer. Similar results have been found in previous studies of other cancers. This makes it unlikely that the cancers will be cured by drugs that target just one or a few genes, the researchers say.

Other scientists call the latest results sobering but important. "Unfortunately this wasn't what we all hoped for," says Stephen Elledge of Brigham and Women's Hospital in Boston, Massachusetts. "But there is useful information in there, and what they're learning is more of what they already learned — which is that cancers are extremely complex."

"What they're learning is more of what they already learned — which is that cancers are extremely complex."

Vogelstein's team analysed nearly 21,000 genes in some 20 patients with a type of brain cancer called glioblastoma multiforme[1] and a similar number with pancreatic cancer[2]. The Cancer Genome Atlas team looked at 600 genes in 91 patients with glioblastoma multiforme[3].

The studies hoped to find genetic glitches driving the tumours that could be treated by drugs similar to Novartis's Gleevec (imatinib) and Genentech's Tarceva (erlotinib), which inhibit the activity of some mutated genes that can cause certain cancers. Instead, the findings confirm earlier hints that patients with the same cancer diagnosis can harbour different sets of genetic causes.



CNRI/SPL

Solid tumours such as glioblastoma (red) can be caused by multiple genes in different patients.

"It is extremely unlikely that drugs that target a single gene, such as Gleevec, will be active against a major fraction of solid tumours," says Vogelstein, whose group has published similar studies on breast and colorectal cancers[4].

The latest papers do identify some single genes that seem to be important in subsets of the cancers. For instance, Vogelstein and his colleagues report that a gene called *IDH1*, which had not previously been linked to brain cancer, was often mutated in younger patients with a certain type of glioblastoma. And the Cancer Genome Atlas group reports that a gene called *NF1*, whose link to cancer had been hypothesized before, was mutated in 23% of 206 patients that were analysed.

The atlas group also found that patients with a particular epigenetic make-up who are treated with one type of chemotherapy show a pattern of genetic and epigenetic changes that may render them resistant to further treatments. The mutation pattern suggests why this resistance evolves and may help doctors find strategies to avoid it.

That shows the importance of the atlas and other similar studies, such as those included in the 10-nation International Cancer Genome Consortium, says Lynda Chin of the Dana-Farber Cancer Institute in Boston, who was the team leader for the atlas paper. "These are very important, clinically relevant questions, and you can't answer them in a traditional hypothesis-driven manner," Chin says.

The Cancer Genome Atlas has moved more slowly than its architects had hoped because it has proved harder than expected to find enough high-quality tissue samples to analyse. The *Nature* paper is an interim analysis of the samples it has studied so far. But most of the genes identified in it and the Vogelstein studies had already been identified in earlier studies, and are not found in most patients with a particular tumour.

That bolsters critics of the atlas, such as Elledge, who has long said that functional studies will be needed to filter the most clinically relevant drug targets out of the massive pool of mutations found in cancer.

"The information they're getting is useful," he says, "but it's expensive and I think some of that money should go to help get you further along into finding drugs." ■

Erika Check Hayden

1. Parsons, D. W. *et al. Science* doi:10.1126/science.1164382 (2008).
2. Jones, S. *et al. Science* doi:10.1126/science.1164368 (2008).
3. The Cancer Genome Atlas Research Network *Nature* doi:10.1038/nature07385 (2008).
4. Sjöblom, T. *et al. Science* **314,** 268–274 (2006).

See Editorial, page 138.

CHEMICAL BIOLOGY
NIH awards $280 million for controversial compound hunt
www.nature.com/news

Genomics institute secures its future

The Broad Institute of MIT and Harvard, a prominent genomics and chemical-biology research centre based in Cambridge, Massachusetts, has received an endowment of US$400 million from philanthropist Eli Broad that sets it on a course to becoming an independent, non-profit organization.

Set up in 2004 as a unique collaboration between the Massachusetts Institute of Technology (MIT) and the Whitehead Institute for Biomedical Research, both also in Cambridge, Harvard University and its affiliated hospitals, the institute now has a more secure future. Earlier gifts — two $100-million donations from Broad — had an expiry date: the funds had to be used within ten years; beyond that, the future of the institute was uncertain.

However, this new gift will be invested and used to support the institute over the long term.

Launched to unite the clinical expertise of Harvard's medical school with academics and engineers at MIT and Harvard, the institute has grown to encompass 1,200 affiliated scientists. It is renowned for its genomics projects, which span the gamut from sequencing the genomes of a menagerie of mammals to looking for genetic factors underlying conditions such as schizophrenia and autism. The centre's chemical-biology programme, meanwhile, screens for small molecules to study and treat diseases. The new endowment will not alter the institute's core research focus, says founding director Eric Lander.

R. FRIEDMAN/CORBIS

The Broad Institute in Cambridge, Massachusetts.

From the start, Broad, a former businessman based in Los Angeles, viewed the institute as an experiment in scientific collaboration. The new donation, announced on 4 September, signals the end of the experimental phase. "We want to see this become permanent, and to have permanence you need to have an endowment," he says.

Faculty members of both MIT and Harvard will continue to hold adjunct positions at either institution, and the two universities will have representatives on the Broad Institute's governing board. "It will continue as an entity of MIT and Harvard," says Lander. "This won't affect how any of the collaborations work." ■

Heidi Ledford

Physicist convicted over links with Iran and China

On 3 September, a federal jury found electrical engineer J. Reece Roth, 70, guilty on 17 counts for conspiracy, wire fraud and violating the Arms Export Control Act. Roth, a professor emeritus at the University of Tennessee in Knoxville, had been working on ways of using plasma to reduce drag on aircraft wings.

He had employed both Chinese and Iranian graduate students without proper authorization (see *Nature* 442, 232; 2006). Roth will be sentenced on 7 January 2009; he faces a maximum sentence of 150 years.

UNIV. TENNESSEE/AP

J. Reece Roth.

The conviction serves as a warning to other academics, says Russ Dedrick, a US attorney with the Eastern District of Tennessee. "Our scientific and educational communities must take precautions to ensure that technology and research are protected," he says.

Lancet retracts paper on stem-cell treatment

The medical journal *The Lancet* retracted an article by urologists at the Medical University of Innsbruck in Austria last week (S. Kleinert and R. Horton *Lancet* 372, 789–790; 2008).

The article claimed positive results for a clinical trial using stem cells to treat urinary incontinence, but this summer an investigation by the Austrian government's Agency for Health and Food Safety found serious flaws in the trial, including incomplete patient consent forms and forged insurance documents (see *Nature* 454, 922–923; 2008). An Austrian Academy of Sciences investigation continues. The university suspended principal investigator Hannes Strasser, but took no sanctions against department head, Georg Bartsch, who was an honorary co-author on the paper. Both have denied wrongdoing.

A *Lancet* editorial accompanying the retraction decried honorary authorships as "unacceptable" and said that such authors still have obligations in cases of flawed research: "With credit comes responsibility — always."

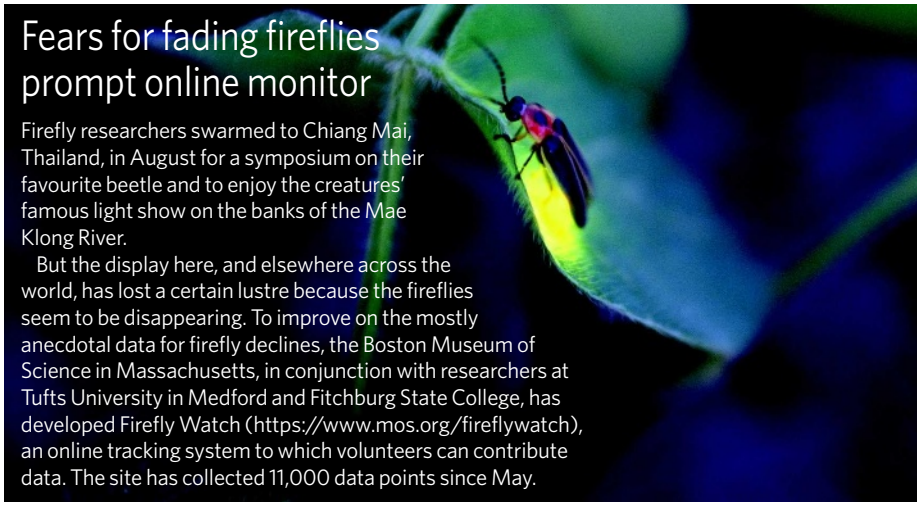
India wins waiver to buy nuclear technology

After three days of deliberation in Vienna, the 45-member Nuclear Suppliers Group (NSG), which regulates global nuclear trade, has bent its rules to allow India to trade with member countries in order to expand Indian civilian nuclear operations.

India has been barred from trading since it first tested nuclear weapons in 1974. The NSG's decision was the result of intense lobbying by the United States — which has a landmark agreement with India pending congressional approval — and support from France and Russia.

The NSG waiver follows a similar endorsement from the International Atomic Energy Agency in July. These actions stand to grant India access to nuclear fuel and technology for energy without its having to sign a non-proliferation agreement. The Communist Party of India called for the Indo-US nuclear deal to be quashed, saying it infringes India's sovereignty.

G. SHUMWAY/GETTY IMAGES



Fears for fading fireflies prompt online monitor

Firefly researchers swarmed to Chiang Mai, Thailand, in August for a symposium on their favourite beetle and to enjoy the creatures' famous light show on the banks of the Mae Klong River.

But the display here, and elsewhere across the world, has lost a certain lustre because the fireflies seem to be disappearing. To improve on the mostly anecdotal data for firefly declines, the Boston Museum of Science in Massachusetts, in conjunction with researchers at Tufts University in Medford and Fitchburg State College, has developed Firefly Watch (https://www.mos.org/fireflywatch), an online tracking system to which volunteers can contribute data. The site has collected 11,000 data points since May.

China pauses on turning coal into liquid fuel

The Chinese government has issued a temporary moratorium on new facilities for converting coal into liquid transportation fuels, although at least two of the biggest projects will still move forward.

The moratorium would allow the state-owned Shenhua Group to build a plant in Inner Mongolia and continue a project with the South African energy giant Sasol. But the broader effects of the policy are unclear for an estimated six other plants reportedly under construction.

Chinese companies have been pursuing coal-converting technologies for a range of products, from chemicals to liquid fuels, to reduce China's reliance on imported crude oil. Industry officials say coal-based chemical production will continue and suggest that the current policy does not necessarily represent a definitive shift away from coal-to-liquid technologies.

Wellcome Trust goes into partnership with India

The Indian government has approved a ten-year, £160-million (US$280-million) partnership between its Department of Biotechnology and the Wellcome Trust, a medical-research charity based in London, UK, to fund biomedical research in India.

Operated as an independent, charitable trust based in New Delhi, the new alliance will award research fellowships to an estimated 700 researchers over ten years. It was announced on 4 September that the two parties will each contribute 50% to the programme.

"The programme will result in a strong, world-class human-resource foundation ... at a scale where the country can make a global impact," according to the cabinet statement.

Maharaj Kishan Bhan, secretary of the biotechnology department, says that it is the most ambitious scheme ever launched by India to reverse the brain drain.

Correction

In the News Feature 'The long summer begins' (*Nature* **454,** 266–269; 2008), the photograph of divers on page 268 should have been credited to Jeremy Stewart not Doug Barber.

Diego Pol with the backbone of a fossil sauropod from central Chubut.

R. DALTON

THE NEW MOTHER LODE

Palaeontologists in Argentina are exploring a trove of fossils that is rewriting evolutionary history. **Rex Dalton** reports.

In the shadow of Cerro Cóndor, a 600-metre-high limestone bluff in Patagonia, two young palaeontologists gaze over waves of mountain ridges running west towards the Andes. Diego Pol and Ignacio Escapa, from the Egidio Feruglio Palaeontological Museum in Trelew, Argentina, have spent years trekking the winding gravel trails here in the Chubut River valley, meeting only wandering guanacos, rheas and sheep. Already, their team has hand-dug half a dozen quarries in nearby canyons that have yielded globally important fossils.

But many prizes remain among the uncharted sediments of the Middle Jurassic, a geological epoch spanning 160 million to 180 million years ago, when dinosaurs, plants and early mammals were all undergoing key evolutionary changes. This time period holds crucial clues to the explosion of evolutionary diversity in both dinosaurs and mammals. The oldest known dinosaur remains, for instance, are around 230 million years old; the oldest known fossil mammals have been dated at 193 million years ago[1]. Both groups diversified to an enormous extent during the Middle Jurassic[2], yet relatively few sediments of that age have been studied. That makes Chubut province in southern Argentina a rare opportunity. "This has the potential to be a global landmark for the Middle Jurassic," says Pol. "For the Southern Hemisphere, it already is."

The Argentine finds may open a little-understood palaeontological window, just as China's rich fossil beds have illuminated the early history of mammals, dinosaurs, reptiles and birds. Chubut is "an amazing region because you get fairly complete skeletal material, which allows you to answer many evolutionary questions", says Peter Makovicky, a palaeontologist at the Field Museum in Chicago who has explored much of Argentina.

"The discoveries from the Middle Jurassic of Argentina are no ordinary field finds," adds Zhe-Xi Luo, a curator at the Carnegie Museum

of Natural History in Pittsburgh, Pennsylvania, who has published on the earliest mammals from China. "They are of such a significant nature that the whole early mammalian evolutionary paradigm must be changed."

Within the past decade, for instance, Argentine fossils have helped rewrite conventional wisdom on the evolution of tribosphenic mammals, so named for having molars of a mortar-and-pestle design that can both grind up plant material and shear meat. Palaeontologists had thought that tribosphenic mammals evolved only on the ancient northern land mass of Laurasia, which included what is now Asia, Europe and North America. But in 2001, within sight of Cerro Cóndor, Pablo Puerta of the Trelew museum unearthed a tiny jaw of a shrew-like tribosphenic mammal, *Asfaltomylos patagonicus*[3]. This confirmed that tribosphenic mammals had evolved on the southern supercontinent, Gondwana, before it began splitting from Laurasia about 180 million years ago. Another team had found one southern Middle Jurassic tribosphenic mammal, in 1999 on Madagascar[4] — but "the Argentine discovery was overwhelming evidence there were multiple evolutions of the same innovation", says Luo.

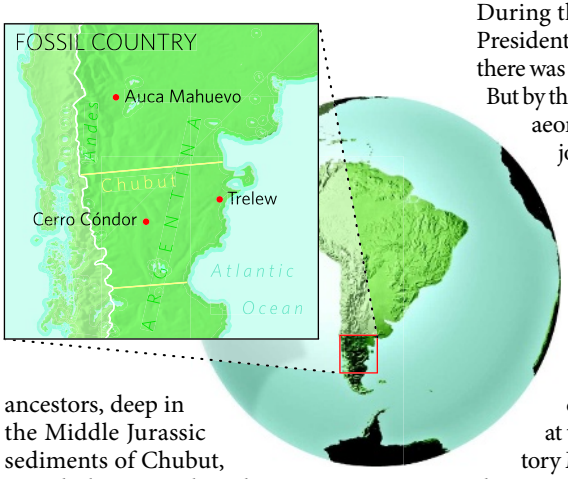
More tribosphenic discoveries may await in the preparatory laboratory at Trelew, where Puerta and his colleagues are close to removing the matrix from about 20 small mammalian specimens. These include parts of skulls and skeletons — rare finds for creatures usually represented by jawbone fossils. "We need to more fully prepare them to know how many new species we have," says Argentine palaeontologist Guillermo Rougier, a mammal specialist at the University of Louisville in Kentucky. "But we will greatly increase what we know of mammals from that time."

Another type of Middle Jurassic mammal has emerged from the Cerro Cóndor quarries — South America's first example of the proto-mammals known as triconodonts, and dubbed *Argentoconodon fariasorum*[5]. In this animal, Rougier and his colleagues see characteristics similar to triconodonts found in North America and Morocco, including teeth like those in modern seals and some other fish-eating mammals.

And last year, Rougier and his colleagues also reported the discovery of a mammal they named *Henosferus molus*, on the basis of three jawbones. Each bone had a lateral groove, marking where cartilage had attached three or four little bones to the back of the jaw[6]. Such bones, which are typically lost during the fossilization process, are the predecessors of the ear bones of later mammals. Rougier believes *A. patagonicus*, *Ar. fariasorum* and *H. molus* are part of the ancestral lineage leading to monotremes — animals that, like the platypus, lay eggs like a reptile but nourish their young with milk.

Dinosaur heaven

Along with its mammals, the Chubut Valley offers new windows on dinosaur evolution. In 2005, Argentine researchers and colleagues in Germany reported the discovery in Late Jurassic rocks of a short-necked sauropod dinosaur[7]. *Brachytrachelopan mesai* had vertebrae that were shorter in length than those of long-necked sauropods. It had clearly evolved to browse on lower-growing plants — and its ancestors, deep in the Middle Jurassic sediments of Chubut, may help to explain how and why.



Much of the time the palaeontologists aren't sure what treasures they have until they get a block of fossils into the lab at the Trelew museum — a decade-old building now a centerpiece of the city's historic downtown, in a frontier village that played host a century ago to outlaws Butch Cassidy and the Sundance Kid. Technicians at the museum are currently removing rock from a large pod of fossils dug up last year. The size of a small convertible, the cache weighs several tonnes. Fossils stick out at various points from its plaster jacket. "We believe it holds a new type of theropod," a two-legged, mainly meat-eating dinosaur, says Pol. "There could be more than one. But we won't know until it is prepared, hopefully by later this year."

> **"This has the potential to be a global landmark for the Middle Jurassic."**
> — Diego Pol

Plant fossils from Chubut are also promising. "I wouldn't be surprised if the first flowering plants come from Middle Jurassic sediments like those in Chubut," says Mark Norell, curator of vertebrate palaeontology at the American Museum of Natural History in New York. Escapa, the palaeobotanist at Trelew, sighs at the thought, saying "that would be great, but I make no predictions". For now, Escapa is happy describing types of cypress never identified in the Southern Hemisphere. Fossils of the new species, called *Austrohamia minuta*, show incredible plant detail, complete with fossilized cones[8].

Despite such potential, Argentine palaeontology has blossomed only in the past decade, due to a national political climate for science that previously varied from deadly to uninterested. When authoritarian generals ruled in the late 1970s, students and scientists were among the tens of thousands of Argentines killed for purportedly having left-leaning political beliefs. During the decade-long term of right-wing President Carlos Menem that ended in 1999, there was little support for any type of research.

But by the early 2000s, papers by Argentine palaeontologists began appearing in major journals, typically from the programme at the Argentine Museum of Natural Sciences in Buenos Aires, run by the now-retired Jose Bonaparte.

A technician without a doctorate, Bonaparte prowled the countryside for four decades making seminal discoveries. His protégés — including Luis Chiappe, now curator of vertebrate palaeontology at the Los Angeles County Natural History Museum, and Rodolfo Coria, former director of the Carmen Funes Museum in Plaza Huincul in Argentina — subsequently discovered dinosaur eggs and revealed dinosaur nesting behaviour at Auca Mahuevo (see map), 100 kilometres or so north of Plaza Huincul. These studies produced detailed analyses of dinosaur eggs, nests and the discovery of a giant predatory dinosaur, from around 80 million years ago[9].

Emerging from the shadows

Around 1980, Bonaparte briefly studied the Cerro Cóndor sediments, which had been discovered in 1949 by Italian-born palaeobotanist Joaquin Frenguelli. But with many other sites available to work, he moved on. Now the Middle Jurassic sediments are coming under scrutiny from researchers such as Puerta, Rougier and a German collaborator, Oliver Rauhut of the Bavarian State Collection for Palaeontology and Geology in Munich. They have been joined by Argentine researchers who returned with doctorates from abroad, like Pol, who completed his Columbia University studies in 2004 at the American Museum of Natural History in New York before coming to Trelew in 2006 as curator of vertebrate palaeontology.

There, he and colleagues have benefited from a 15-year-old Argentine policy to establish museums in provinces to exhibit and study specimens. A coastal town founded by Welsh immigrants nearly 125 years ago, Trelew — Welsh for 'town of Lewis', named for its founder — is a prominent stop for tourists heading to Patagonia and beyond. With about 110,000 visitors a year bringing in revenue, the museum has developed a healthy research programme. Its palaeontological exhibit area will soon be doubled, at cost of US$5 million. "They have the best laboratory equipment in Argentina," says Makovicky.

Local residents know it. Unlike in China, where farmers often scavenge fossil sites for specimens for sale[10], digging in Argentina is controlled by national policy and private landowners. Near Cerro Cóndor, in a village of a half-dozen homes, the community has embraced its palaeontological history, setting up a mini-museum to educate children and naming its traditional annual fiesta as the dinosaur holiday.

When Pol and Escapa visit, locals show them fossils they have found when riding after their livestock. But more often than not, the most promising new localities come from the palaeontologists' own kilometres of hiking. As they survey the barren ground, Escapa looks for dark or black rocks — indicating carbonaceous material that was sealed ages ago, limiting oxygen so that a plant fossil can form. The first finds are often conifers, such as the cypress. Once a promising fossil is noted, the hikers stop, break rocks and dig. Frog and amphibian fossils usually are the first indication of vertebrates. With luck, those fossils lead to larger vertebrates.

On a brisk autumn day in June this year, Pol and Escapa checked such a site, where they had earlier found a metre-long dinosaur specimen, including the brain case, a valued find for any new species. They had hoped to be able to extract the fossil that day, but logistical problems caused by ash drifting from a Chilean volcano delayed its removal.

Rocks at the site also contain conifer and other plant fossils.

R. DALTON

Extracting fossils can be onerous. For the large pod back at the lab, it took five years from the day in 2002 when technician Leandro Canessa discovered the fossils jutting from the hillside. First, the fossil clump had to be covered with casting material; then arrangements were made with a construction crew to bring a bulldozer to cut a road up a steep canyon. Finally, a crane was driven in to winch the fossil cluster into a truck. "This was really a project," says Pol, giving credit to Puerta. But the hard work pays off in providing the exact geological context for a fossil. This stands in stark contrast to China, where palaeontologists often have to reconstruct the geology of a purloined fossil long after it has been removed from the ground.

With each new find, the Cerro Cóndor scientists divide up the research based on their specialities. Beginning in 2000, Rauhut worked on dinosaurs, with Pol joining him later. Escapa takes the plants.

There's no shortage of work to go around. The Middle Jurassic, says Rauhut, is "the least-known part of dinosaur history. And the area around Cerro Cóndor is incredibly rich." Since first coming as a postdoc to Trelew in 2000, Rauhut has been on 13 field campaigns. New finds that have yet to be described include a sauropod and an ornithischian dinosaur, a bird-hipped creature. "The ornithischia is the one I am excited about," says Pol. "We have a skull, a lower jaw and about 50% of the skeleton. There is nothing known about Middle Jurassic ornithischia from this region."

And perhaps the biggest challenge is just the processing time to analyse and study all the new fossils. With a new three-year grant of €90,000 (US$130,000), Rauhut and his colleagues will be heading out to Cerro Cóndor again later this year, hoping to bring as many specimens back as possible from the field. "There is a lot more to come," he says. ■

Rex Dalton is a reporter for *Nature* based in San Diego.

The Cerro Cóndor bluff overlooks a treasure trove of fossils.

R. DALTON

1. Luo, Z. X., Crompton, A. W. & Sun, A. L. *Science* **292,** 1535–1540 (2001).
2. Luo, Z. X. *Nature* **450,** 1011–1019 (2007).
3. Rauhut, O. W., Martin, T., Ortiz-Jaureguizar, E. & Puerta, P. *Nature* **416,** 165–168 (2002).
4. Flynn, J. J., Parrish, J. M., Rakotosamimanana, B., Simpson, W. F. & Wyss, A. R. *Nature* **401,** 57–60 (1999).
5. Rougier, G. W. *et al. Am. Mus. Novitates* No. 3580, 1–16 (2007).
6. Rougier, G. W., Forasiepi, A. M., Martinelli, A. G. & Novacek, M. J. *Am. Mus. Novitates* No. 3566, 1–54 (2007).
7. Rauhut, O. W., Remes, K., Fechner, R., Cladera, G. & Puerta, P. *Nature* **435,** 670–672 (2005).
8. Escapa, I., Cueno, R. & Axsmith, A. *Rev. Palaeobot. Palynol.* doi:10.1016/j.revpalbo.2008.03.002 (2008).
9. Chiappe, L. M. *et al. Nature* **396,** 258–261 (1998).
10. Dalton, R. *Nature* **406,** 930–932 (2000).

NATURE|Vol 455|11 September 2008

The race to break the standard model

The Large Hadron Collider is the latest attempt to move fundamental physics past the frustratingly successful 'standard model'. But it is not the only way to do it. **Geoff Brumfiel** surveys the contenders attempting to capture the prize before the collider gets up to speed.

It is powerful; it is galling; it is doomed. The incredibly successful mathematical machine that physicists call the 'standard model' is a set of equations that describes every known form of matter, from individual atoms to the farthest galaxies. It describes three of the four fundamental forces in nature: the strong, weak and electromagnetic interactions. It predicts the outcome of one experiment after another with unprecedented accuracy. And yet, as powerful as it is, the standard model is far from perfect. Its mathematical structure is arbitrary. It is littered with numerical constants that seem equally ad hoc. And perhaps most disturbingly, it has resisted every attempt to incorporate the last fundamental force: gravity.

So physicists have been trying to get beyond the standard model ever since it was put together in the 1970s. In effect, they will have to shatter the model with experimental data that contradict its near-perfect equations. And then, from its fragments, they must build a newer, better theory. The Large Hadron Collider (LHC), a giant particle accelerator at CERN, Europe's particle-physics laboratory near Geneva, Switzerland, is the latest attempt to break the standard model — and one that many see as all but assured of success. The prodigious energy it generates will force particles into realms where the standard model cannot follow. In the race to move beyond the status quo, "the LHC is by far the favourite", says Frank Wilczek, a theorist at the Massachusetts Institute of Technology in Cambridge who won the 2004 Nobel Prize in Physics for his work underpinning the standard model.

But the LHC is not the only game in town. For decades physicists have tried to get beyond the standard model in all sorts of ways, sometimes with accelerators, sometimes with precision measurements of breathtakingly rare events, sometimes with observation of outer space. And in the time it takes for the LHC to get fully up to speed — its first results aren't expected until at least next summer (see 'The unstoppable collider') — some of those experimental groups think that they have a fighting chance of seizing the prize first. Their task will be hard: the standard model is a formidable piece of work that has resisted all the easy and obvious attacks. To crack it, experiments will need unprecedented sensitivity, a multitude of data, and more than a little luck.

ILLUSTRATIONS BY J. RIORDAN

Here's a rundown of the heroic few who feel up to the task.

TEVATRON

While the LHC gets its protons up to speed, the world's other heavyweight particle-accelerator is racing to break the standard model first. Since 2001, the Tevatron, located at Fermilab in Batavia, Illinois, has been accelerating protons and antiprotons at an energy of around 1 tera electron volt.

That's only a seventh of the eventual top energy of the LHC, but total energy isn't everything in the hunt for new physics. Collisions that would generate new particles outside the standard model are extremely rare, which means that the longer an accelerator runs and the more data it accumulates, the better its chances of finding something. So for a while, at least, the Tevatron will continue to have a data lead over the LHC. Even by the summer of 2009, the Tevatron will have several times more total data than its new competitor.

And already those data are showing some tantalizing, if tentative, hints of something beyond the standard model. One deviation comes in measurements of a particle known as the strange B (B_s) meson. The B_s is made of a strange quark and an anti-bottom quark, and it is among the heaviest of all mesons. Under a rule known as charge-parity symmetry, the standard model predicts the B_s will decay in the same way as its antiparticle (made of an anti-strange and a bottom quark). But measurements of the two are hinting at a difference in their decays. According to Dmitri Denisov, a spokesperson for the D-Zero experiment at the Tevatron, that difference could be an important clue in the quest for discoveries. It might signal the existence of new, exotic particles, or of previously unknown principles. In any case, says Denisov, "it's an exciting measurement".

The B_s anomaly is not the only oddity showing up at the accelerator, adds Robert Roser, a spokesperson for the Tevatron's other major experiment, the Collision Detector at Fermilab, or CDF. An unusual feature in the decays of pairs of top and anti-top quarks has him intrigued. Again, he admits, it's far from certain. But some of these signals may turn out to be important, Roser says. "As you add data, one of [these anomalies] may become real."

Perhaps not surprisingly, a more sceptical view comes from John Ellis, a theorist at CERN. Yes, the Tevatron could provide some tantalizing hints, says Ellis, but it is unlikely to make a definitive find before the LHC comes on strong. In the world of particle physics, he points out, nothing constitutes a discovery until it is measured to five σ (five standard deviations from the mean), the equivalent of 99.99994267% accuracy. Much more data than the Tevatron has accumulated so far will be needed to reach that exacting standard, and the detector is unlikely to make those big gains before it is overtaken by its new rival. "I think its going to be very, very tough for the Tevatron," Ellis says. "I just don't see them getting it before the LHC starts going gangbusters."

COSMOS

While the high-energy physicists gather in their machine's control rooms, another group of physicists is looking to the heavens. There they hope to find something that shatters the standard model — if the Universe cooperates.

The main thing that their spacecraft will look for are indications of dark matter, the ghostly substance that could make up as much as 85% of the matter in the Universe. Astronomers know that dark matter exists only because of its gravitational pull on galaxies and its influence on the Universe's shape; it seems to pass right through the kind of ordinary matter found in stars, planets and people. Presumably, dark matter is a haze of particles that rarely, if ever, react with the ordinary variety. But nobody is quite sure what those particles might be — except that they are not accounted for in the standard model.

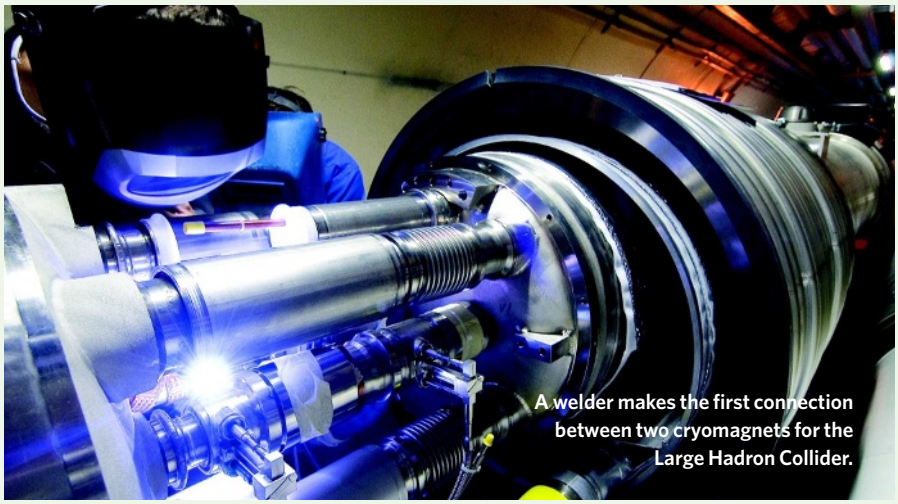
One candidate comes from the 'supersymmetry' theory, which predicts that every particle in the standard model has another, heavier partner lying outside the model. The lightest of these supersymmetric partners is called the neutralino, and is predicted to have just the right properties to be dark matter.

Neutralinos themselves wouldn't be seen by telescopes, orbiting or otherwise. But periodically, two neutralinos could collide and annihilate — creating a shower of more mundane particles that orbiting detectors might pick up. The PAMELA (Payload for Antimatter Matter Exploration and Light-nuclei Astrophysics) experiment has already seen an intriguing clue. The satellite-borne instrument has unofficially reported a surplus of anti-electrons that may have been generated by dark-matter annihilations (see *Nature* **454,** 808; 2008). "It's a beautiful result," says Graciela Gelmini, a physicist at the University of California, Los Angeles, who has seen PAMELA's data. Still, she adds, the complexities of the measurement require caution.

A second, recently launched satellite may also be able to spot the untimely demise of the neutralino. The Fermi Gamma-ray Space Telescope is a US$690-million space instrument designed to scan the entire sky for ultra-high-energy photons. It is possible that such γ-rays could be created by neutralino collisions, in which case they would show up as a ubiquitous haze in the orbiting detector's sky-map. "That would be a stunning, stunning signature," says Steven Ritz, the telescope's project scientist at NASA's Goddard Space Flight Centre in Greenbelt, Maryland.

Such signatures, if they're spotted and confirmed in time, definitely have a chance to beat the LHC in the quest to break the standard model, says Michael Turner, a cosmologist at the University of Chicago in Illinois. But Ritz points out that although astrophysics could technically be the first to make such a discovery, they can't do much more than that. Anti-electrons, γ-rays and other such signatures could provide physicists with only a rough mass range for the new particles, and would say nothing about how supersymmetry might work. For those reasons "there would still be a large number of essential question marks", says Ritz — questions that would have to be resolved at the LHC.

M. BRICE/CERN



A welder makes the first connection between two cryomagnets for the Large Hadron Collider.

The unstoppable collider

As *Nature* went to press, the Large Hadron Collider (LHC) at CERN, Europe's particle-physics laboratory near Geneva, was on the verge of circulating its first protons. But there is much to be done before the machine produces publishable scientific findings. In the coming months, even as operators fine-tune the collider itself, other physicists will be trying to get the experiments spaced around the ring up and running.

Switching on a detector the size of a building is no small task. Each instrument is made of hundreds of thousands of smaller detectors, which must be synchronized to track the particles generated by collisions. The detectors are currently being brought into alignment using cosmic rays from outer space, says Peter Jenni, the spokesperson for the ATLAS (Toroidal LHC Apparatus) experiment. But watching particles from real collisions will be a different matter entirely. The colliding proton beams will produce hundreds of millions of distinct 'events' every second, each event comprising hundreds or thousands of debris particles flying outward from the collision point. As the detectors are designed to track most or all of these particles individually, the result will be far more data than the experimentalists can handle. Fortunately, the vast majority of the collisions will produce nothing out of the ordinary. So the experimenters have equipped their detectors with electronic 'triggers' that separate the interesting collisions from the rest. For example, one simple trigger will tag collisions that produce 'muons' — particles that can be created by the decay of more massive particles. Each trigger will be designed to save the evidence of a certain kind of interesting event, and each must be carefully tuned, according to Jenni.

After the data are filtered, they must be analysed. To that end, data from the experiments will be sent to thousands of physicists via a massive computing grid that can shuttle petabytes of data to university labs around the globe. Initial trials have gone well, says Jim Virdee, the spokesperson for the Compact Muon Solenoid (CMS) experiment at CERN, another major experiment, and teams at ATLAS and the CMS are now drilling with computer-generated practice data.

Assuming everything goes smoothly, Jenni and Virdee both say that results could come as early as summer of 2009. By then, the accelerator should have been running for a few months at its full 7-tera-electron-volt strength, and there will have been time to sort out any technical issues.

Will the LHC find some new physics in that first run? Possibly. The machine will collide particles at roughly seven times the energy of the world's current leading accelerator, the Tevatron, located at Fermilab in Batavia, Illinois. That's a big jump, and it will, in principle, be possible to see new particles almost immediately, says Virdee. "You don't need that much data to probe beyond what Fermilab has done," he says.

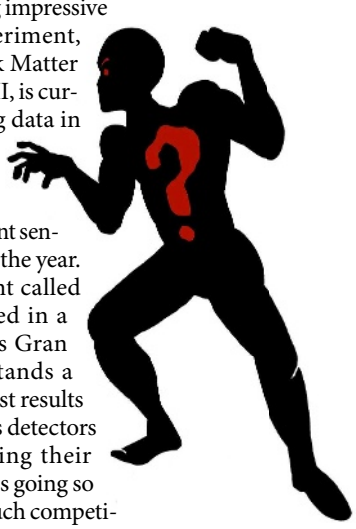
Fermilab physicists are understandably sceptical of that assessment. It took two full years for physicists working at the Tevatron to fully grasp the idiosyncrasies of their experiments, says Robert Roser, a spokesperson for the Collision Detector at Fermilab. And even with the higher energies, it will take a significant number of collisions to find something new, adds Dmitri Denisov, a spokesperson at Fermilab's D-Zero experiment. "Colliding one proton with one proton at the centre of a detector will not be enough," he says. **G.B**

THE DARK

Other physicists have chosen darkness over light. From their lairs inside disused mines and traffic tunnels, they are watching a number of highly sensitive detectors that could find direct signatures of dark matter, including supersymmetric neutralinos (see *Nature* **448,** 240; 2007).

There are around half-a-dozen different schemes for such detectors, but they all follow the same basic concept. Take some stuff you think could respond to dark matter, place it deep underground to protect it from cosmic rays and other disruptive influences, and wait for something to happen. "It's like watching grass grow," says Wilczek.

Although they are perhaps not the most exciting way to beat the LHC, these detectors are making impressive progress. One experiment, the Cryogenic Dark Matter Search II, or CDMS II, is currently accumulating data in the Soudan Mine deep beneath Minnesota. Its operators aim to treble its current sensitivity by the end of the year. Another experiment called XENON100, located in a tunnel under Italy's Gran Sasso Mountain, stands a chance to have its first results out before the LHC's detectors can finish processing their findings. "The field is going so fast and there's so much competition, that it's not easy to survive at the moment," says Elena Aprile, the principal investigator for XENON100 at Columbia University in New York. "It's an amazing time."



And on top of these prospects, one group claims that it has already seen dark matter in its detector. Earlier this year, the DAMA/LIBRA (Dark Matter Large Sodium Iodide Bulk for Rare Processes) experiment, also at the Gran Sasso National Laboratory, announced that it had seen a signal in its latest generation of detector (see *Nature* **452,** 918; 2008). But their finding has the other groups stumped, says Aprile, whose experiment sits in a vault next to that of DAMA/LIBRA. No one else has yet been able to confirm the signal, and in fact, the findings from other teams seem contradictory, she says. "We are definitely not consistent."

Although these detectors seem to be improving in leaps and bounds, they have an Achilles heel: they only work if the so-far unseen dark-matter particles interact, at least occasionally, with regular matter. There's no guarantee that that is the case, says Ellis. And

as far as he's concerned, that makes these experiments "shots in the dark".

Still, Ellis concedes that there is a chance that these esoteric searches might manage to see something before the LHC can. "I think the dark-matter guys are the jokers in the pack," he says.

NEUTRINO

The next few months will be a caffeine-fuelled blur for most of those scientists racing to beat the LHC. But neutrino physicists can take it easy: they've already broken new ground, and they did it a decade ago.

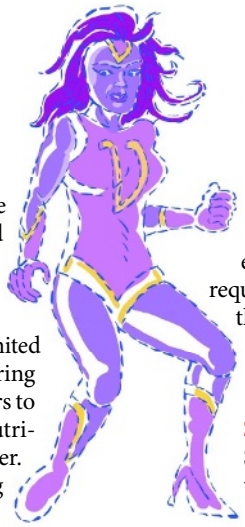
Neutrinos are the neutral members of the 'lepton' family of particles, the group that includes the electron. The original version of the standard model predicted that neutrinos should be completely massless, but experimentalists suspected otherwise. For years they saw fewer neutrinos from the Sun than theorists predicted. One possible explanation for the deficit was that solar neutrinos could be switching from one type to another. But that switching would be possible only if neutrinos had mass. In 1998, a Japanese experiment in Hida called Super-Kamiokande saw the neutrino switch in action, and that result is the first — and to date the only — firm finding that defies the standard model.

Unfortunately, says Ellis, the neutrino's mass can be accommodated within the standard model by making just a few simple modifications to the equations. "It's possible to add something in relatively easily," he says. And consequently, although neutrino physicists can arguably claim the prize, their discovery hasn't helped theorists in their search for new models of physics.

But neutrinos may not be finished just yet. Experiments in the United States, Europe and Japan are now firing beams of neutrinos at their detectors to try to learn more about how the neutrinos switch from one kind to another. The precise details of this switching may help narrow the field of possible new theoretical models, says Lisa Randall, a theorist at Harvard University in Cambridge, Massachusetts.

And two new detectors could go further still. A European collaboration is now running the Astronomy with a Neutrino Telescope and Abyss Environmental Research (ANTARES) detector under the Mediterranean Sea off the coast of Toulon, France, and a team of Americans is installing IceCube beneath the ice of Antarctica. Both use strings of detectors to see high-energy neutrinos from cosmic sources striking water or ice. ANTARES was completed earlier this summer, whereas IceCube has about half of its 70 strings of detectors installed. But already IceCube is five times more sensitive than Super-Kamiokande, according to Francis Halzen, IceCube's principal investigator at the University of Wisconsin, Madison. "It's not inconceivable we'll find something," he says.

Just what that something might be is up for debate. One possibility would be neutrinos produced by dark-matter particles trapped in the Sun's core. But again, Halzen says, anything seen by the neutrino experiments would almost certainly require follow-up by the LHC. "I think these experiments are complementary," he says. "But if you give me a choice, I'd rather see it first."



SUCCESS?

So can any of these projects best the standard model? Wilczek is sceptical. "I'm not on the edge of my seat," he says. Looking the track record, it seems that, "the standard model always wins". He believes that only the LHC stands a real chance of breaking the existing paradigm.

And there's no guarantee that even the giant collider will find something new. "Supersymmetry could show up anytime between mid-2009 and never," says Ellis. If never is the date, he says, physicists will face "the maximum conceivable horror scenario". "What will we do next?" he asks.

But Turner takes a different view. Ultimately, these experiments and the LHC are fighting the battle together. He is confident that by combining their data with the LHC's, the standard model can be bested, and that new physics will be discovered. "We're on the verge of a major revolution," he says. ■

Geoff Brumfiel is a senior reporter for *Nature* based in London.

To read more about the LHC start-up, visit the *Nature News* special at http://tinyurl.com/5usrfl.

NATURE|Vol 455|11 September 2008

THE FATE OF FINGERS

Proteins with 'zinc fingers' designed to bind almost any DNA sequence will soon be available to any lab that wants them — from two very different sources. **Helen Pearson** reports on a revolution in designer biology.

"There is only one word that matters in biology," says pioneering molecular biologist Aaron Klug, "and that is specificity. The truth is in the details, not the broad sweeps." This neatly explains the importance of proteins equipped with structures called zinc fingers, on which Klug's team at the MRC Laboratory of Molecular Biology in Cambridge, UK, did pioneering work in the 1980s and 90s. The human body contains more than 700 different zinc-finger proteins, which bind to specific DNA sequences to switch genes on and off. For almost 20 years, would-be protein engineers have dreamed of taking the DNA-recognition ability of these zinc fingers and making it universal — of designing zinc-finger proteins targeted at any DNA sequence that catches their fancy. The world is about to reap the benefits of a decade of hard work realizing those dreams.

> **"How can we get this in the hands of every lab in the world that wants to use it?"**
> — Edward Lanphier

From the middle of September, designer zinc-finger proteins will be available to anyone with an idea, an Internet connection and US$25,000. The reagents company Sigma-Aldrich, based in St Louis, Missouri, has a splashy launch planned for CompoZr, a service that will provide its customers with zinc-finger proteins aimed at whatever DNA sequences they want. The service uses techniques developed by Sangamo Biosciences of Richmond, California. Zinc-finger technology has previously been the preserve of a select few, mostly working with Sangamo; with Sigma's cut-and-paste offering, Sangamo hopes to make zinc fingers far more widely used in commercial research and in academia.

The specificity zinc fingers offer could find a million uses in the lab, revolutionizing the techniques researchers use to work out what genes do. This summer two studies, one from Sangamo[1] and one from Scot Wolfe's lab[2] at the University of Massachusetts, Worcester, showed that bespoke zinc-finger proteins that can cut the DNA they recognize — zinc-finger nucleases — could, in principle, knock out any gene in zebrafish. This is something that researchers working with most model organisms have so far been unable to do with other methods. In an accompanying article[3], Ian Woods and Alexander Schier of Harvard University wrote that such tools might become "the major technology for genome manipulation". "It could have a monstrous impact," says Wolfe.

The technology could also prove valuable in the clinic. Monoclonal antibodies, a previous breakthrough in biological specificity, went through a phase of being just lab tools. Now they are the basis of a therapeutic market worth over $20 billion, and treat tens of thousands of patients annually. Sangamo hopes that therapies that use zinc fingers to turn genes on and off — or indeed to edit them — have similar potential. Earlier this year, in an illustration of what might be possible, the company built a zinc-finger protein that disables a protein by means of which HIV gains access to human cells. When applied in mice, the zinc-finger protein locked the virus out[4].

Exploiting that sort of capability, Sangamo plans first to open up and then dominate a whole new class of therapeutics. On the wall of his office looking towards San Francisco Bay, chief executive Edward Lanphier has framed copies of inspirational headlines from biotech history — such as the *San Francisco Examiner*'s 1980 "Genentech Jolts Wall St", commemorating the stock-market debut of the company on the other side of the bay that has done as much as any to capitalize on the specificity of antibodies. Asked whether he entertains such visions, though, Lanphier demurs: "My hubris is significant, but not that significant."

The CompoZr offering is intended in part to appease researchers who have expressed frustration at not being able to access Sangamo's proprietary technology. But following developments a couple of months ago, those academics may end up spoilt for choice. In July, a powerful new methodology for do-it-yourself zinc-finger

design was published in *Molecular Cell* by a consortium of scientists led by Keith Joung of the Massachusetts General Hospital in Boston[5]. Joung hopes the consortium's protocol — the culmination of ten year's work on his part — will allow any lab, anywhere, to make zinc-finger proteins at a fraction of Sigma's price.

Two ways forward

Joung says that he and the consortium are not seeking to undercut Sangamo, but rather to expand the zinc-finger universe. Still, some will inevitably see the two as rivals — or even as a new round in the struggle between academics devoted to open systems and companies built on the defence of intellectual property. "It's pretty unusual to have an academic research group as committed to an open platform in clear opposition to a firm," says Arti Rai, an expert in patent law and the biopharmaceutical industry at Duke University in Durham, North Carolina, who has studied Sangamo. Michael Eisen, at the Lawrence Berkeley National Laboratory in California, is blunter: "It's nice there's a reservoir of rebellion and people saying 'screw it, we're not going to be held back'."

Although the first zinc-finger protein was discovered in 1985 (ref. 6), the protein-engineering possibilities only really came to the fore in a May 1991 paper in *Science*[7] by Carl Pabo and Nikola Pavletich, then both at Johns Hopkins University in Baltimore, Maryland. This paper revealed the X-ray structure of three zinc-finger domains bound to a piece of DNA. The zinc fingers — each a chain of 30 amino acids folded back on itself and stabilized by a zinc ion — nestle in the major groove of the DNA molecule, touching three of the base pairs that provide the DNA with its sequence. The fingers lie one after the other in the groove, with three fingers recognizing a sequence of nine bases overall. "Everyone who saw that structure had the same thought," says Joung. "It might make a very nice scaffold for making designer DNA-binding proteins." Understand the relationship between the amino acids and the bases and you might design a protein that recognized any sequence of bases you chose, threading zinc fingers together as simply as beads on a string.

> **"By a very aggressive, creative and smart licensing strategy Sangamo has swept up most of the intellectual property."**
> — Arti Rai

Lanphier, though, saw more than an engineering opportunity in the fingers. He saw a business break. Lanphier had spent his career in the business and strategic side of the pharmaceutical and biotechnology industries. In the 1990s he was head of commercial development at a California company called Somatix Therapy, which was developing the vectors that deliver genes into tissues for gene therapy. But although the company controlled intellectual property (IP) surrounding its vectors, others owned the genes that the company might want to put into them. "We were frustrated with the fact that we couldn't access proprietary genes," Lanphier says.

Zinc-finger proteins offered the opportunity to create those genes — newly written genes to describe newly imagined proteins. "If you have a motif that can be engineered, by definition those different proteins will be encoded by different genes," Lanphier says. "So you might have a technology platform here to generate an infinite number of genes." One of those designer genes, delivered to a cell, would make a zinc-finger protein that could control one of the cell's naturally occurring genes. With that in mind, Lanphier flew out to talk to some of the leading academic groups in the field, including Pabo's and Klug's.

Bolstered by those meetings and half a million dollars in venture capital, Lanphier started Sangamo in 1995. He hardly looks the part of a cut-throat monopolist as he pads around his office in hiking socks and Birkenstocks, but he went about developing the company with a fierce focus: he wanted an unassailable IP position. "By a very aggressive, creative and smart licensing strategy they've swept up most of the IP," Rai says. "It's unusual to have this vision."

Both inside and outside Sangamo, early

attempts to engineer zinc-finger proteins were based on two attractive and sometimes complementary ideas. One was the *a priori* approach: work out the rules specifying which amino acids direct a finger to which of DNA's 'letters' (the bases A, G, C and T) and rationally design new fingers. The other was the 'look-for-what-works' approach: design or try to identify zinc-finger domains that bind to each of the 64 possible three-letter sequences, and then mix and match them like Lego blocks to fit whichever base sequence is wanted.

If only the science of life were so simple. Researchers soon realized that there was no simple code, and that 'modular assembly' was not as easy as it looked. For one thing, the sequences bound by a series of zinc fingers turned out not to be completely distinct from each other but instead overlapped: the same DNA triplet can be touched by amino acids from more than one finger (see 'In the groove'). Almost everyone in the field came to believe that each finger had to be chosen in the context of the fingers next door. "You have to take into account the overlap or the failure rate is very high," says Mark Isalan of the EMBL/CRG Systems Biology Research Unit in Barcelona, Spain, who worked in Klug's lab and devised a way to design proteins that took context into account[8]. He took the technology forward at Gendaq, a company co-founded by Klug.

In April 2000 — amid a frenzy of anticipation over the human genome sequence — Sangamo took its stock public, raising $50 million in a day when the exuberant market valued the company at $375 million. In 2001, it bought Gendaq, and Isalan's approach has formed the basis of the company's protein building ever since. Flush with new money and science, Sangamo started to hit its stride.

Mix and match

Joung first got to grips with zinc fingers in 1998, when he joined Pabo's lab at the Massachusetts Institute of Technology (MIT) in Cambridge as a postdoc inspired by the idea of designing proteins *de novo*. Then — as now — he set about the task by a process of selection: "evolution in a tube", he says. Joung builds libraries of zinc-finger combinations, gets them mass produced in bacteria, and then identifies those that are best at binding a specific sequence. The strength of this approach is that if you start off with combinations of zinc fingers, rather than looking at them one at a time, the issue of context gets dealt with. The problem is that to cover all the possible combinations of key amino acids in a three-finger protein you would have to make 8×10^{24} different proteins. Joung calculates that working with such a library in bacteria would require an agar plate around 100 times the size of North America. By the end of his postdoc his library-building and selection method worked — but only for finding a single finger, not a set[9].

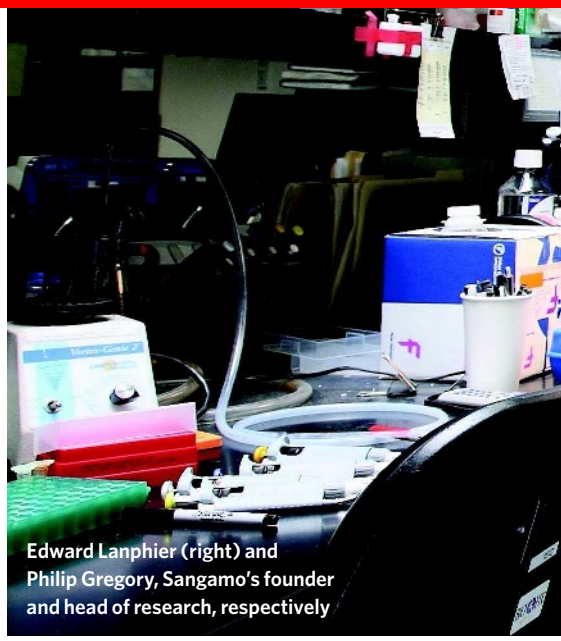
In September 2001, Pabo told his lab that he was moving to California to work for Sangamo. Joung says there was some discussion of him moving to Sangamo, too — the company had assessed his technology and licensed it, although it didn't end up using it — but he wanted an academic position, and one became available at Massachusetts General Hospital. Isalan, as it happens, made a similar decision: "[Sangamo] offered me a job but I'm really a scientist, and by that stage it was a production line," he says.

In the development of that production line, the company has built an extensive library of mainly two-finger proteins, designed so as to take into account interactions between neighbours. When the company's scientists want to design a protein for a new target gene, they feed the gene sequence into a computer program, which tells them various ways to piece together two-finger modules into a four-finger protein that might serve, and predicts how well each option will work. A robot then assembles the DNA for the new zinc fingers, picking out genetic fragments encoding the various modules from a standardized set of 384-well plates. "We've spent a lot of time and energy developing this kit of pieces, information about how these pieces function and then the software that knows best how to use them," says Jeff Miller, a one-time lab-mate of Joung's who made the transition from MIT to Sangamo with Pabo.

And the amount of time and energy Miller and his colleagues spent on honing their approach, not to mention the $230 million the company spent on this and other R&D, is far beyond what an academic group could hope to repeat. Another company might, but although no one can really know the strength of an IP position until it is tested in court, Sangamo's seems to have been strong enough to scare off most competition.

It's possible that this is good for the company but bad for the field. Richard Jefferson, founder of CAMBIA, a non-profit organization based in Canberra, Australia, that supports open access to life-sciences technology, claims that if a company had pursued monoclonal antibody or DNA sequencing IP as diligently as Sangamo has pursued zinc-finger IP "it would have set [those fields] back a long way". Lanphier rejects the idea: "On the contrary, I would argue that our patents have allowed us to access the funding that we have used to advance zinc-finger technology."

Sangamo has not just been amassing IP and developing its design processes: it has also started showing what zinc fingers might do. Much of the recent excitement in the field has focused on the zinc-finger nucleases, originally discovered in 1996 (ref. 10). In 2005, working with Matthew Porteous of

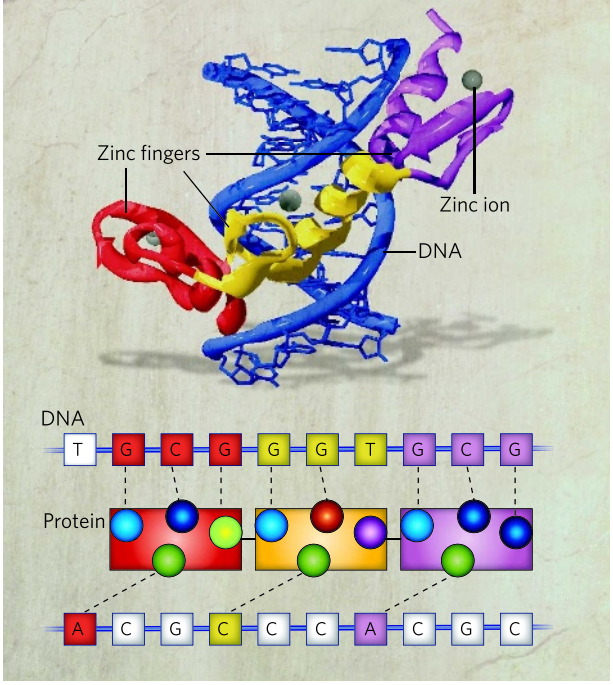


Edward Lanphier (right) and Philip Gregory, Sangamo's founder and head of research, respectively

E. MILLETTE

IN THE GROOVE

Three zinc fingers of the protein Zif268 nestle alongside the bases in their target stretch of DNA (top). Amino acids in each finger associate with specific bases in the DNA (bottom).



SOURCE: JAMIESON, A. C. *ET AL. NATURE REV. DRUG DISC.* **2**, 361–368 (2003)

the University of Texas Southwestern Medical Center in Dallas, scientists at Sangamo showed that a zinc-finger nuclease aimed at a mutation in a gene called *IL2Rγ* erased the mutation in over 18% of cells, converting the gene into one that worked properly[11]. The mutation in question causes X-linked severe combined immune deficiency (SCID), a fatal genetic disease; the results looked all the more promising because gene-therapy trials for SCID a few years previously — in which a working copy of the gene was introduced into cells — had been successful in compensating for the defect but resulted in several children developing leukaemia. "People were very excited," Porteous says. The efficiency was good enough that if it could be made to work *in vivo* it might actually cure people.

Spreading the message

Sangamo has been racking up further scientific and commercial testaments to the potential applications of mutation-correcting zinc-finger nucleases. Plant biologists are drooling with delight over the enzymes, because they should enable precise genetic modifications to be made in plants to genes that have been impossible to target precisely with conventional techniques — opening up new possibilities for adding to, knocking out or overwriting genes associated with yield, nutritional content, pest resistance and other useful traits.

Zinc-finger believer: Keith Joung.

The company has an agreement with Dow Agrosciences, based in Indianapolis, Indiana, to develop the technology for plants. "It's revolutionizing plant breeding — that's not too strong a word for it," says Klug, who serves on Sangamo's scientific advisory board. Together with drug giant Pfizer and Genentech, Sangamo is also engineering cell lines to improve their protein productivity. And the company has an agreement with Genentech's corporate parent Roche for the creation of transgenic cells and animals.

But as Sangamo went from strength to strength, progress in the field as a whole was not so inspiring. Only a handful of groups, such as Joung's and Wolfe's, were able to make proteins using their own methods. "Many people have got excited by the technology and try it and it doesn't work and they give up," Isalan says. "It's been a real problem." Sangamo thus became a target of envy. "I think there has been some frustration," says Porteous, who no longer collaborates with the company. "They publish and present at conferences as if they've invented sliced bread, but they are very closed about helping anyone else do it."

> **"People get excited by the technology and try it and it doesn't work and they give up."**
> — Mark Isalan

Joung — who, were he not devoted to the lab, would make a first-rate diplomat — is less critical: "I think Sangamo has done a really amazing job of moving the field forward." The problem he saw was not that Sangamo had too much technical ability — it was that the rest of the world had too little. "For many years I'd felt academics needed to develop a technology of their own," he says. "Sangamo was putting so much time and resources into this, there was no way that one lab was going to be able to keep up with them, so if we were to stay relevant to the field we had to band together."

In the summer of 2005, Joung got in touch with another zinc-finger veteran, Dan Voytas, then at Iowa State University. In 2000, Voytas had started a small company called Phytodyne to develop zinc-finger nucleases for plants. Voytas says that the company folded in 2004 after failing to come to a deal with Sangamo over access to the company's proteins. "We were bumped out of business. Their pockets had more cash in them than ours." (Sangamo says it would have loved to collaborate, but Phytodyne's lack of funding made it too unstable to partner with.) The zinc-finger consortium that grew out of a dinner Voytas and Joung had later that year now has 14 member labs, including many big names in the field, but most of its protein engineering has been done in Joung's lab.

Joung's current technique uses a battery of small libraries, or what he calls 'pools', that each contains just less than 100 proteins selected to bind a target sequence. One pool, for example, is full of protein domains that bind well to the sequence GAA, but only when they are the first finger in a three-finger protein. Another pool contains domains that bind to GAA when they are the second finger, and so on. To create a three-finger protein that binds at a specific sequence, three pools are combined and subjected to a second round of selection, from which the best-binding proteins are pulled out. "This is an advanced form of mix and match," Joung says. Sixty-eight of 192 possible pools now sit behind the ice-caked door of a −80 °C freezer off his lab.

Joung doesn't have a robot — but two years ago he hired a young technician called Morgan Maeder and she has been spending near-robotic hours honing the technique. The first time she showed that the pools had generated a zinc-finger nuclease that corrected a gene in human cells "was the most exciting thing ever",

A. BEKKER

A. BEKKER

Production factors: Morgan Maeder in Joung's lab (top); Jeff Miller and Sangamo's robot (right)

she says. "Even Keith was giddy; he said 'give me a high five' and we were jumping up and down, it was so exciting. For me it's two years' worth of work — for him it's a whole career." On 24 July, the consortium published Joung's methods for zinc-finger protein engineering[5], along with work from other members of the consortium showing that the proteins accurately cut genes in human and plant cells. They named the technique OPEN, for Oligomerized Pool Engineering. Joung is now arranging to make the pools available to other labs; he estimates the set will cost around $5,000. But the technique is hardly workaday. Maeder says her protocol runs to 20 pages.

Porteous says he would have liked to be a fly on the wall in Richmond when Lanphier's crew first saw the OPEN paper. Whatever such a fly might have heard, though, in public Sangamo has no problem. The company is much more focused on its first clinical results. Later this year, phase II trials will reveal whether a zinc-finger protein designed to activate a growth factor called VEGFA has promoted the recovery of damaged nerves in diabetics: the first chance for zinc fingers to show beneficial effects in humans. And in the next few months the company also plans to file two investigational new drug applications with the US Food and Drug Administration to test two zinc-finger nucleases — including one that disrupts the HIV receptor CCR5 — in humans for the first time.

Philip Gregory, Sangamo's vice-president of research, does not think that OPEN will be much competition for the click-and-go convenience of ordering from Sigma. "If you're a biologist and want to knock out a gene you might not want to set up as a zinc-finger lab to answer that one question. So with labs less fiscally constrained the advantage of working with a quicker product, and what's probably a better product anyway, will be convincing."

But convenience does not always win out. Eisen sees parallels with the stand-off in the 1990s between Affymetrix of San Jose, California, and academics, including Eisen, who thought the cost of the company's DNA chips was restricting research and published their own system for manufacturing microarrays that did the same thing. "Just the mindset — the transition from thinking it's inaccessible to something people could use — was important," he says. "It allowed arrays to be integrated into the fabric of research in a way that they wouldn't have if they had remained a commercial product only, and I can see exactly the same here."

An integrated, improvable, hands-on experience might give OPEN an edge in some academic settings. But that could suit Sangamo fine. Success for the consortium will increase the total amount of zinc-finger research, and that may matter more to Sangamo than what reagents are being used. Lanphier says that one of the original motivations for the CompoZr service came from Sangamo's management asking themselves why biotech companies focused on RNA interference were gaining value while theirs wasn't. They concluded that RNAi was a hot technology in part because it was widespread. "RNAi was ubiquitously available and easy to use, high-school kids were using it," recalls Lanphier. "So we said: 'how can we get this in the hands of every lab in the country or the world that wants to use it?'." The more zinc-finger research is done, in this analysis, the more exciting Sangamo looks.

> **"All the consortium is doing is loading up Sangamo's pipeline."**
> — Richard Jefferson

And the more commercial possibilities it might have. Sangamo is confident that when that research gets to the stage of generating products, its hard-bought IP position will win through. "Ultimately, from a commercialization perspective, you'd run into Sangamo's IP at some point," says Gregory. If Sangamo's patent position is as strong as the company claims, then "all the consortium is doing is loading up Sangamo's pipeline", says Richard Jefferson.

Since the OPEN paper came out, Joung says he has had some 15 requests for the pools. Now that researchers have the proteins in their hands, he says, they can start to tackle all kinds of other pressing questions, such as how to deliver them efficiently to cells, and whether they might make unwanted cuts in the genome. "Unless you were willing to use [Sangamo's] proteins there was nothing to work with — now we can ask those questions," says Joung.

But for all the possible applications, his personal fascination remains the motif itself. "I look at the sequences as a way to see that they worked, but I don't really care what they are," says Maeder. "But he loves it. Every time it works, it's like further proof that the past ten years of his career have been worthwhile." Joung and some of his colleagues still think it's possible that there is some kind of code — an algorithm for writing the best amino-acid sequence with which to target a piece of DNA from first principles. It's even conceivable that the tools with which to crack that code are sitting in the databases that are now filling up with successful designs, awaiting release through bioinformatic wizardry.

Still, complex though it is, there's something satisfying about the current process — at least to Joung. "You start with 200 million different things and you end up with this very small number that are very similar," he says. "It's exciting and gratifying to me that the system actually works and that you don't always get the same finger depending on which context it was selected in. I had the thought nearly ten years ago, so yes, it's pretty cool to see it come out this way." To come out in all its glorious specificity — which is what matters in biology. ■

Helen Pearson is *Nature*'s biology features editor.

1. Doyon, Y. *et al. Nature Biotechnol.* **26,** 702–708 (2008).
2. Meng, X., Noyes, M. B., Zhu, L. J., Lawson, N. D. & Wolfe, S. A. *Nature Biotechnol.* **26,** 695–701 (2008).
3. Woods, I. G. & Schier, A. F. *Nature Biotechnol.* **26,** 650–651 (2008).
4. Perez, E. E. *et al. Nature Biotechnol.* **26,** 808–816 (2008).
5. Maeder, M. L. *et al. Mol. Cell* **31,** 294–301 (2008).
6. Miller, J., McLachlan, A. D. & Klug, A. *EMBO J.* **4,** 1609–1614 (1985).
7. Pavletich, N .P. & Pabo, C. O. *Science* **252,** 809–817 (1991).
8. Isalan, M., Klug, A. & Choo, Y. *Nature Biotechnol.* **19,** 656–660 (2001).
9. Joung, J. K., Ramm, E. I. & Pabo, C. O. *Proc. Natl Acad. Sci. USA* **97,** 7382–7387 (2000).
10. Kim, Y. G., Cha, J. & Chandrasegaran, S. *Proc. Natl Acad. Sci. USA* **93,** 1156–1160 (1996).
11. Urnov, F. D. *et al. Nature* **435,** 646–651 (2005).

CORRESPONDENCE

These letters respond to the Commentary 'The science of doping' by Donald A. Berry (*Nature* **454,** 692–693; 2008).

Doping: a paradigm shift has taken place in testing

SIR — Donald Berry claims that anti-doping tests are based on flawed statistics. Your Editorial 'A level playing field?' (*Nature* **454,** 667; 2008) goes even further in concluding that the anti-doping authorities act unscientifically. These claims neglect an abundant body of literature and ignore the paradigm shift that has taken place in anti-doping science.

Anti-doping is a forensic science, not a medical one. In medical diagnostics, biostatisticians have all the leeway to set sensitivity and specificity to an appropriate level. Such freedom is restricted in forensics: the risk of a false positive must be minimized at every step of the development, validation and application of a test. This fact alone explains why anti-doping tests do not necessarily rely on statistical reasoning, and certainly not solely on a specificity threshold, something Berry seemingly takes for granted. For the detection of exogenous testosterone in particular, anti-doping laboratories establish intervals for a reference population throughout validation processes that also include quality controls for batch acceptance. To date, no false positive has been reported among all the negative controls.

The nature of scientific evidence is also different. In forensics, the traditional assumptions of 'absolute certainty' and 'discernible uniqueness' are being progressively abandoned in favour of an empirical and probabilistic approach (see M. J. Saks and J. J. Koehler *Science* **309,** 892–895; 2005). In the fight against doping, this is embodied by the 'athlete's biological passport', an electronic document that stores an individual's information pertaining to indirect markers of doping.

In multiplying the probabilities to estimate the specificity for the Landis case, Berry makes a basic statistical error. Indeed, successive tests are not independent in a longitudinal follow-up (P. E. Sottas *et al. Forensic Sci. Int.* **174,** 166–172; 2008).

A more thorough literature search would have prevented Berry from attempting to reinvent the wheel and from concluding that anti-doping scientists are "on the wrong path", which is presumptuous and disrespectful. The role of anti-doping science (not "doping science") is to protect clean athletes. Your Editorial may have just the opposite effect.

Pierre-Edouard Sottas, Christophe Saudan, Martial Saugy
Swiss Laboratory for Doping Analyses, Chemin des Croisettes 22, 1066 Epalinges, Switzerland
e-mail: pierre-edouard.sottas@chuv.ch

Doping: probability that testing doesn't tell us anything new

SIR — In his Commentary, Donald Berry discusses Bayes' rule, noting that consideration of *P*, the prior probability of guilt, is essential in interpreting a positive doping result. He fails, however, to mention what the actual value of *P* might be in Floyd Landis's case, which I think misses an opportunity to address an important problem.

Athlete acquaintances and the news media have led me to believe that *P* can be very high, and in fact approach unity, in some sports. If this is true, then anti-doping measures should cease — and not because of the statistical arguments that Berry raises, rather because the testing isn't telling us anything we don't already know.

If *P* is close to 1, then negative tests are most likely to be false negatives. Those who test positive might only be those who are least adept at hiding their drug use.

Geoffrey Baird Department of Laboratory Medicine, Division of Clinical Chemistry, University of Washington, Seattle, Washington 98195, USA
e-mail: gbaird@u.washington.edu

P. DEJONG/AP

Disqualified Tour de France winner Floyd Landis still asserts his innocence.

Doping: ignorance of basic statistics is all too common

SIR — Donald Berry's Commentary is like a breath of fresh air in the murky world of drug testing. Unfortunately, a lack of competence in basic statistics is all too common in biology and the clinical sciences. As Berry points out, there is often a lack of accounting for pre-test probabilities in the application of tests with known sensitivities and specificities, as well as for issues arising from multiple testing.

Even those who grasp the principles of Bayes' rule frequently make the mistake of not empirically confirming the utility of confirmatory assays. Take steroid testing, as illustrated in Berry's Figure 1 for Floyd Landis's case in 2006. Given the high sensitivity and specificity of the assay, androsterone plus 5α-androstanediol is assumed to form the basis of a conclusive set of tests for confirming positive screening results with etiocholanone plus 5β-androstanediol. In fact, the confirmatory tests can provide little additional information unless they have been shown to be independent predictors of drug positivity.

Matthew Fero Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA
e-mail: mfero@fhcrc.org

"Oppenheimer told me of a problem that was very much on his mind."
François de Rose on the birth of CERN, page 174

Doping: similar problems arise in medical clinics

SIR — Donald Berry warns about the dangers of poor statistical understanding and misinterpretation of drug-testing results in Olympic athletes. Unfortunately, this same problem arises on a daily basis around the world in medical clinics, often with even greater consequences.

Berry illuminates the failure to use proper Bayesian reasoning in interpreting doping tests and also the problem of not having sufficient control-population norms for the tests to determine correctly whether an athlete is taking a banned substance or not. Clinicians typically have less understanding of Bayesian statistics than drug-testing officials and even fewer resources to interpret or norm such tests.

Take urine testing of patients on opiate therapy to make sure that they test positive for opiates (to show the patient is taking the medicine rather than, say, selling it) and that they are not using illegal drugs. Either a negative test for opiates or a positive test for an illegal substance can typically be sufficient to preclude a patient from receiving another prescription for opiates or to put the clinician in the position of having to explain the test result before prescribing the medicine.

Such tests need to be reported with the appropriate Bayesian interpretation. Also, as Berry advocates for Olympic athletes, patients should have the right (and access) to a statistical 'consultation' if they feel the test is in error.

Eric L. Altschuler Department of Physical Medicine and Rehabilitation, University of Medicine & Dentistry of New Jersey, Newark, New Jersey 07103, USA
e-mail: altschel@umdnj.edu

Playing the system to give low-impact journal more clout

SIR — A hundred years or so ago, a patent officer who was bored with his routine work wrote up his speculations on light quanta (A. Einstein *Ann. Phys.* **17,** 132–148; 1905), citing other people's work to avoid long explanations. Today, there is a whole citation industry that — among other things — affects the impact factors of scientific journals, which in turn provide a gauge for the quality of an institution's research output.

Publication in prestigious journals that have high impact factors encourages researchers to pursue trendy topics. It follows that investigators working in low-profile and under-researched fields are at a disadvantage because they publish in less well-known journals and generate fewer citations. This not only offends their institutions; in countries such as the Czech Republic it could fail to stimulate the flow of grant money.

The Swiss journal *Folia Phoniatrica et Logopaedica* has a good reputation among voice researchers but, with an impact factor of 0.655 in 2007, publication in it was unlikely to bring honour or grant money to the authors' institutions.

Now two investigators, one Dutch and one Czech, have taken on the system and fought back. They published a paper called 'Reaction of *Folia Phoniatrica et Logopaedica* on the current trend of impact factor measures' (H. K. Schutte and J. G. Švec *Folia Phoniatr. Logo.* **59,** 281–285; 2007). This cited all the papers published in the journal in the previous two years. As 'impact factor' is defined as the number of citations to articles in a journal in the past two years, divided by the total number of papers published in that journal over the same period, their strategy dramatically increased *Folia*'s impact factor this year to 1.439.

In the 'rehabilitation' category, shared with 26 other journals, *Folia* jumped from position 22 to position 13. Publication there will now no longer disappoint the Dutch author's colleagues for lowering their institution's score, and should encourage the Czech government to spend more money on the Czech author's university.

Could professional scientometrists one day be in demand, to guide young scientists up the citation ladder of scientific survival and allow them to do some good, modest science in their spare time, just for fun?

Tomáš Opatrný Faculty of Science, Palacký University, Svobody 26, 77146 Olomouc, Czech Republic
e-mail: opatrny@optics.upol.cz

Changing the rules won't stop the rise of a new superpower

SIR — In their Essay 'The end of the science superpowers' (*Nature* **454,** 412–413; 2008), J. R. Hollingsworth and colleagues argue that the pattern of rise and decline of science superpowers such as France, Germany and Britain is now catching up with the United States. Surprisingly, they see a shift not towards Chinese scientific hegemony but towards multipolarity.

They argue that the decline of the United States indicates the end of a model of scientific production, that 'big' science is finished and that small interdisciplinary institutes, where new ideas can flourish, are taking over. In this context of altered dynamics, they conclude that US science can prosper alongside contributions from elsewhere.

This argument ignores a persistent pattern in the history of science. Calls for interdisciplinarity and creativity always arise when leaders are confronted with new competition from outside. Such calls are often a sign that the callers are losing this competition.

One of the strengths of science is that its rules of engagement are clear, making it possible for anyone to participate if they take the effort to learn the rules. This means that there is always room for newcomers taking scientific development to its next logical step, overtaking formerly dominant elites. As the authors point out, this happened to France in the mid-nineteenth century, to Germany in the 1920s, and to Britain after the Second World War.

A typical reaction of elites under threat is to raise entry barriers to their circle by placing emphasis on knowledge unavailable to newcomers. For example, they may trade the universal language of their science for methods relying on culturally specific 'general knowledge' and interdisciplinary meta-perspectives that come only with a broad education. Laborious scientific methods no longer suffice; creativity and reflection count. For outsiders, the road to success by acquiring leadership in specialized fields is blocked.

History teaches us that discourses of interdisciplinarity and creativity offer temporary refuge for embattled elites, but eventually do not stop the process of shifting scientific hegemony. They result in isolated, inward-looking scientific communities. Much of the post-hegemonic academic discourse in France and Germany illustrates this. If the United States is to avoid this fate, it should increase scientific funding rather than trying to shield itself from competition by changing the rules.

Robbert Maseland Radboud University Nijmegen/Max Planck Institut für Gesellschaftsforschung, Paulstrasse 3, 50676 Cologne, Germany
e-mail: r.maseland@fm.ru.nl

Contributions may be submitted to correspondence@nature.com.

BOOKS & ARTS

There's no place like home?

A bold attempt to synthesize the effects of geography on the world's population through maps highlights some interesting paradoxes, explains **Yi-Fu Tuan**.

The Power of Place: Geography, Destiny, and Globalization's Rough Landscape
by Harm de Blij
Oxford University Press: 2008. 304 pp.
$27.95, £14.99

P.U. EKPEI/AFP/GETTY IMAGES

Sub-Saharan Africa is harsh for inhabitants but has a richer cultural diversity than Europe.

For those who want to be on top of world events, yet feel overwhelmed by the amount of information that floods into their homes and offices through the media, *The Power of Place* is an excellent start. In his new book, professor, writer and broadcaster Harm de Blij uses the geographer's favourite tool, the map, to help us feel somewhat in control of this mass of information — the resurgence of religious fundamentalism, the levelling of the playing field for the well educated, the roughening of the landscape for the illiterate and poor, the threats of climate change and of nuclear and biological terrorism.

He depicts global population distribution, groupings of language families, dominant belief systems, recent earthquake centres, places of recurrent conflict, and other phenomena, at world, regional and local scales. De Blij also puts us at ease by subsuming the materials under a few overarching themes.

A prominent theme is the uneven distribution of wealth: the difference between rich and poor, north and south, core and periphery. These terms are seldom defined by those who use them. North and south are particularly vague, and de Blij doesn't give a definition either. But he does draw a clear line between core and periphery: the core of wealth extends sinuously from Europe and North America to coastal China and the Jakota triangle of Japan, South Korea and Taiwan, and then dips south to Australia and New Zealand. His map also shows places where fences, walls and other devices have been set up to prevent or limit members of the periphery from entering the core.

With the map's help, de Blij articulates a major problem of our time: the core needs a labour force and the periphery needs jobs and income. This relationship would seem to call for arrangements of mutual benefit, yet it is rich in paradox and irony. It collides with the core's desire to defend its 'high' culture — its ideals of democracy, equal rights, individualism, a secularist world view and future orientation — against dilution by large inflows of the poor from the periphery, with their attachment to place, kinsfolk and religion. There was a time when newcomers gradually and willingly merged into the mainstream: the outstanding example in the late nineteenth and early twentieth centuries was the United States. But in the past 30 years, immigrants have been less willing to merge because they see their language and culture as their identity, and as a source of pride not to be given up. Rather than adapt to the core, newcomers believe that the core should adapt to them; for example, that core inhabitants in the United States should learn Spanish and respect male domination in the household. Ironically, the liberal cosmopolitan core has encouraged this trend by promoting the acceptance of ethnic and cultural diversity. With this pluralism comes the idea that one culture is as good as another, only different. If this is true, it begs the question of why we should make the effort to change and move up the cultural ladder. Furthermore, what do we mean by 'up'?

Another major theme is that of the book's title — the power of place. De Blij does not define what he means by power, choosing to let the reader deduce its meanings from the contexts in which he uses it. This is not helpful. Power could mean empowerment, or the opposite — deprivation of autonomy through confinement and control. Place obviously empowers by providing resources for its inhabitants to survive. It does not dictate how far the inhabitants will progress by using the resources. Some inhabitants go far and, over time, produce the high living standard of the core. But resource, as the late geographer Carl Sauer said, is a term of cultural appraisal. Culture determines what we consider a resource and how to make the best use of it. For people to prosper, both place and culture have to empower. Of the two, culture is by far the more important enabler.

De Blij gives many examples of disablement by natural forces such as earthquakes, tsunamis, floods, droughts, tropical storms and endemic diseases. People are at a disadvantage when they live in places prone to natural disaster. They are also at a disadvantage when their geographical location is unfavourable — land locked, for example, or remote from the lanes of commercial and cultural exchange. People of the periphery suffer from these disabling powers of place. But they suffer even more from the peculiar forms of culture they developed in that place. Culture can become a

handicap, discouraging people from enriching themselves and developing further. Many people in the periphery bear the burden of culture even more than the burdens of nature and natural habitat. Culture may be a home for them, but easily turns into a cosy prison.

As de Blij and others have pointed out, the world's poor are concentrated in the tropical latitudes. Here we have another paradox. The warm tropical latitudes are exceptionally rich in flora and fauna; by comparison, the deserts and middle latitudes are poor in plant and animal life. The same paradox seems to apply to culture. Take language and religion, for example; New Guinea has some 900 languages and sub-Saharan Africa has around 2,000; Europe, by contrast, is home to only about 200 languages. The same disproportion is true of religion. Peoples in the tropics have many polytheistic belief systems in which they worship countless spirits and deities, and assign divine powers even to animals, plants and rocks. Peoples of the deserts and steppes, on the other hand, tend to be monotheistic, their belief systems simple and austere.

When we, members of the core, think of plants and animals, we always consider diversity to be a good thing and do our best to preserve it. This preference is sometimes carried over to human languages and cultures. Thus, like many other linguists of the core, de Blij laments the decline in the number of languages in tropical latitudes, forgetting that in New Guinea and Nigeria, the multiplicity of tongues is a barrier to the broad exchange of goods and ideas that is necessary for progress.

The Power of Place is full of fascinating facts, such as this one that I chose at random: global migration, large as it is, makes up less than 3% of the world's population. Despite de Blij's attempts, the mind still finds it hard to make sense of so much disparate information. He should have offered fewer facts, made a greater effort to subsume them under three or four linked concepts, and drawn simpler and stronger conclusions. ■

Yi-Fu Tuan is emeritus professor of geography at the University of Wisconsin-Madison, 550 North Park Street, Madison, Wisconsin 53706, USA. His most recent book is *Human Goodness*.
e-mail: ytuan@geography.wisc.edu

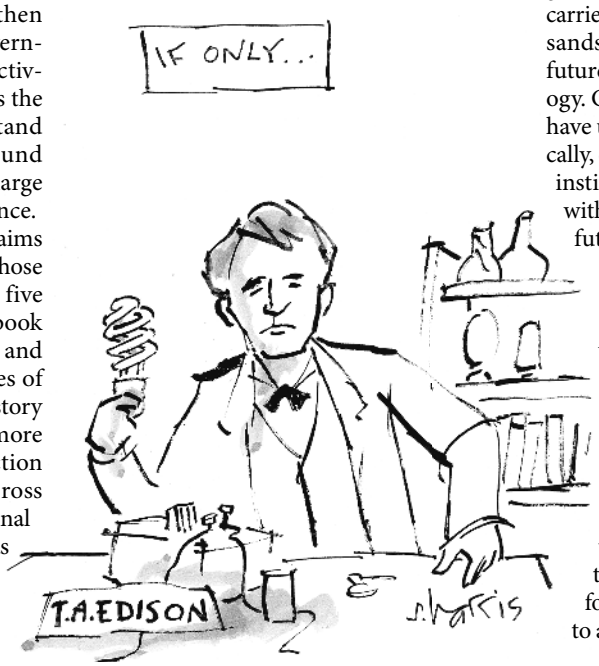
A toolbox for policy planners

The Handbook of Technology Foresight
Edited by Luke Georghiou, Jennifer Cassingena Harper, Michael Keenan, Ian Miles and Rafael Popper
Edward Elgar: 2008. 456 pp. £115, $220

During the past decade, many national governments have sponsored formal planning processes called technology foresight. Such exercises involve a wide range of stakeholders in anticipating long-term social, economic and technological developments, and then using the resulting vision to inform government policies. The growth of foresight activity, most prominently in Europe, reflects the desire of these governments to understand and influence today's rapid and profound social and economic changes, driven in large part by advances in technology and science.

The Handbook of Technology Foresight aims to shape this emerging field and to assist those planning foresight activities. Edited by five scholars active in foresight practice, the book opens with a critical review that defines and distinguishes foresight from other types of futures studies, alongside an excellent history of the field and a detailed summary of more than 30 methodologies. The second section surveys national foresight activities across Europe, Asia and the Americas, and the final section addresses common themes such as evaluation and policy transfer.

Initially a means of informing government investment priorities for research and development, the process of national technology foresight has expanded to address a full range of societal issues that affect and are affected by science and technology. The authors quote approvingly the definition of foresight given by the European Commission's FOREN project, which describes it as "a systematic, participatory, future-intelligence gathering and medium-to-long-term vision-building process aimed at present-day decisions and mobilizing joint actions". They name three characteristics that distinguish technology foresight from other approaches to futures studies. It looks to the future; it uses information about the future to inform near-term decisions; and it includes a broad range of individuals in group exercises to develop forecasts and explore their policy implications.



Foresight in hindsight. (From *101 Funny Things About Global Warming* by Sidney Harris and Colleagues; Bloomsbury, 2008.)

The book's survey of national programmes demonstrates that foresight activities are shaped by the particular needs, culture and politics of a country. The United Kingdom's foresight programme was established in 1993 and has become an institutionalized policy instrument for many agencies and departments. It uses a wide variety of methods such as scenarios, simulations and gaming, workshops and the Delphi interactive expert-based survey for forecasting. By contrast, the Japanese government's technology foresight programme, which has been running since 1969, carries out a nationwide Delphi survey of thousands of experts every five years to map out future developments in science and technology. Central and Eastern European countries have used technology foresight only sporadically, often hindered by political mindsets and institutional structures that are more at ease with single rather than multiple views of the future, and with wholly separate government research endeavours rather than integrated national innovation systems.

Those considering a foresight exercise will find this book a valuable compendium that offers lessons to be learnt, and help in choosing goals, selecting methods and identifying successes and failures. Scholars will find a rich survey of current practice, methodological approaches and tensions in the field. But the book does not address the fundamental question of when national technology foresight can provide an appropriate means to achieve a society's goals.

Technology foresight aims to create a 'national public good'. At a time of fast-paced radical change, it seeks to offer a

handicap, discouraging people from enriching themselves and developing further. Many people in the periphery bear the burden of culture even more than the burdens of nature and natural habitat. Culture may be a home for them, but easily turns into a cosy prison.

As de Blij and others have pointed out, the world's poor are concentrated in the tropical latitudes. Here we have another paradox. The warm tropical latitudes are exceptionally rich in flora and fauna; by comparison, the deserts and middle latitudes are poor in plant and animal life. The same paradox seems to apply to culture. Take language and religion, for example; New Guinea has some 900 languages and sub-Saharan Africa has around 2,000; Europe, by contrast, is home to only about 200 languages. The same disproportion is true of religion. Peoples in the tropics have many polytheistic belief systems in which they worship countless spirits and deities, and assign divine powers even to animals, plants and rocks. Peoples of the deserts and steppes, on the other hand, tend to be monotheistic, their belief systems simple and austere.

When we, members of the core, think of plants and animals, we always consider diversity to be a good thing and do our best to preserve it. This preference is sometimes carried over to human languages and cultures. Thus, like many other linguists of the core, de Blij laments the decline in the number of languages in tropical latitudes, forgetting that in New Guinea and Nigeria, the multiplicity of tongues is a barrier to the broad exchange of goods and ideas that is necessary for progress.

The Power of Place is full of fascinating facts, such as this one that I chose at random: global migration, large as it is, makes up less than 3% of the world's population. Despite de Blij's attempts, the mind still finds it hard to make sense of so much disparate information. He should have offered fewer facts, made a greater effort to subsume them under three or four linked concepts, and drawn simpler and stronger conclusions. ■

Yi-Fu Tuan is emeritus professor of geography at the University of Wisconsin-Madison, 550 North Park Street, Madison, Wisconsin 53706, USA. His most recent book is *Human Goodness*.
e-mail: ytuan@geography.wisc.edu

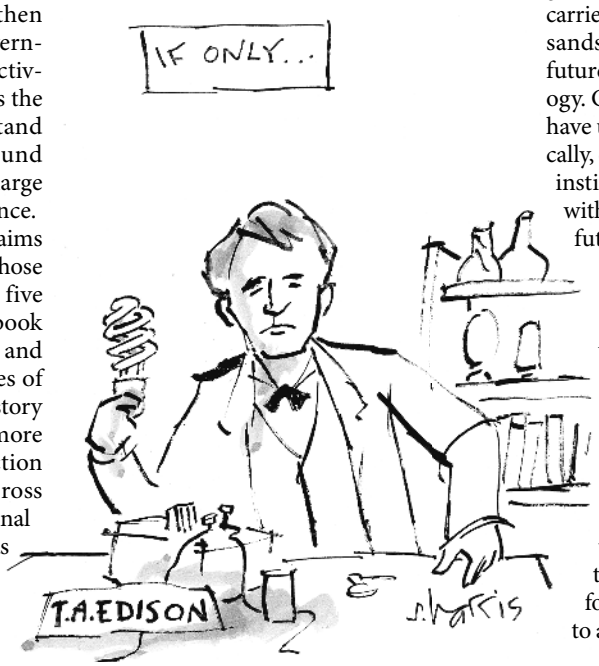
A toolbox for policy planners

The Handbook of Technology Foresight
Edited by Luke Georghiou, Jennifer Cassingena Harper, Michael Keenan, Ian Miles and Rafael Popper
Edward Elgar: 2008. 456 pp. £115, $220

During the past decade, many national governments have sponsored formal planning processes called technology foresight. Such exercises involve a wide range of stakeholders in anticipating long-term social, economic and technological developments, and then using the resulting vision to inform government policies. The growth of foresight activity, most prominently in Europe, reflects the desire of these governments to understand and influence today's rapid and profound social and economic changes, driven in large part by advances in technology and science.

The Handbook of Technology Foresight aims to shape this emerging field and to assist those planning foresight activities. Edited by five scholars active in foresight practice, the book opens with a critical review that defines and distinguishes foresight from other types of futures studies, alongside an excellent history of the field and a detailed summary of more than 30 methodologies. The second section surveys national foresight activities across Europe, Asia and the Americas, and the final section addresses common themes such as evaluation and policy transfer.

Initially a means of informing government investment priorities for research and development, the process of national technology foresight has expanded to address a full range of societal issues that affect and are affected by science and technology. The authors quote approvingly the definition of foresight given by the European Commission's FOREN project, which describes it as "a systematic, participatory, future-intelligence gathering and medium-to-long-term vision-building process aimed at present-day decisions and mobilizing joint actions". They name three characteristics that distinguish technology foresight from other approaches to futures studies. It looks to the future; it uses information about the future to inform near-term decisions; and it includes a broad range of individuals in group exercises to develop forecasts and explore their policy implications.



Foresight in hindsight. (From *101 Funny Things About Global Warming* by Sidney Harris and Colleagues; Bloomsbury, 2008.)

The book's survey of national programmes demonstrates that foresight activities are shaped by the particular needs, culture and politics of a country. The United Kingdom's foresight programme was established in 1993 and has become an institutionalized policy instrument for many agencies and departments. It uses a wide variety of methods such as scenarios, simulations and gaming, workshops and the Delphi interactive expert-based survey for forecasting. By contrast, the Japanese government's technology foresight programme, which has been running since 1969, carries out a nationwide Delphi survey of thousands of experts every five years to map out future developments in science and technology. Central and Eastern European countries have used technology foresight only sporadically, often hindered by political mindsets and institutional structures that are more at ease with single rather than multiple views of the future, and with wholly separate government research endeavours rather than integrated national innovation systems.

Those considering a foresight exercise will find this book a valuable compendium that offers lessons to be learnt, and help in choosing goals, selecting methods and identifying successes and failures. Scholars will find a rich survey of current practice, methodological approaches and tensions in the field. But the book does not address the fundamental question of when national technology foresight can provide an appropriate means to achieve a society's goals.

Technology foresight aims to create a 'national public good'. At a time of fast-paced radical change, it seeks to offer a

commonly shared vision of the future and to create new networks that enable a society to invest its science and technology resources more wisely, harness the beneficial effects of innovation, and ameliorate its risks. Yet foresight may provide only one way to create these benefits. As the book describes, the French government sponsors a small number of comprehensive foresight activities. By contrast, in the United States, many groups — from the independent but government-funded National Academy of Sciences to numerous non-profit organizations — offer visions of the future and build networks around them. Clearly, these approaches reflect different political and cultural contexts. But the different visions may provide different strengths and weaknesses, for instance, offering coherent actions versus resilience to surprise. The book helps frame questions about, for example, which approaches governments should choose, but it does not answer such questions.

To call this volume a handbook may be premature. The word connotes an easily consulted reference that provides quick answers to those engaged in an activity. As the editors note, technology foresight remains a diverse and experimental practice whose theoretical foundations are poorly understood and whose successes have not yet moved from the anecdotal to the empirically grounded. Much remains to be learned. Meanwhile, *The Handbook of Technology Foresight* provides an important survey of current knowledge that will help governments use foresight to navigate these tumultuous times. ■

Robert Lempert is director of the RAND Frederick S. Pardee Center for Longer Range Global Policy and the Future Human Condition, 1776 Main Street, Santa Monica, California 90401, USA.
e-mail: lempert@rand.org

Potatoes and poverty

Propitious Esculent: The Potato in World History
by John Reader
William Heinemann: 2008. 315 pp. £18.99

Propitious Esculent is not just a book about potatoes; it is also about poverty. The two are linked by history, and in this very readable account, anthropologist and journalist John Reader shows us how.

The cultivated potato, *Solanum tuberosum*, is one of around 1,500 species in the flowering plant genus *Solanum*, which also includes the tomato, aubergine and woody nightshade. There are some 190 species of wild potatoes, all found in the Andes — from these a single species has been domesticated and spread throughout the world. Those who only see potatoes in heaps at supermarkets may be surprised that the crop comes from a flowering plant and is one of South America's greatest contributions to the European diet. The United Nations Food and Agriculture Organization has declared 2008 the International Year of the Potato to raise awareness of its importance.

Botanically, the potato is a tuber, a swollen piece of underground stem where the plant stores starch. Wild potato plants and local 'primitive' varieties have tubers, but they are often small and oddly shaped, bearing little physical resemblance to those from cultivated varieties. Reader cites recent research on the taxonomy and domestication of these varied and complicated plants. Disappointingly, he does not incorporate recent work on the genetics of the potato and its relatives.

Nor does the book sufficiently discuss the genetic modification of potatoes for control of disease, an important issue for food poverty. Potatoes are one of the most expensive food crops in terms of pest and disease control. Reader cites the potato as the "world's most chemically dependent crop — with the global cost of fungicides standing at [US]$2 billion per year". This astounding figure comes almost as a footnote at the end of a long chapter about the discovery of *Phytophthora infestans*, the agent of potato blight. The fungus, now associated with the devastation of world potato crops, was first discovered in grapevines. Reader's account of the disease, its discovery and its action is riveting, but potatoes are almost incidental to his story. Today, we easily see the connection between symptom and disease and can then search for causative agents: this was not so in the days when people thought microorganisms arose from spontaneous generation.

Reader does detail the development of blight-resistant potato varieties through the plant-breeding work of Redcliffe N. Salaman at the University of Cambridge, UK, in the early twentieth century. But he does not discuss more modern and controversial approaches. Disease control is being developed at the International Potato Center in Peru through 'true potato seed' potatoes — the crop is replanted using the seed from the original potato plant rather than vegetatively propagated from small pieces of tuber. These varieties have great promise for improving the gene pool for disease resistance, especially in the Andes, where genetically engineered potatoes cannot be used because of their potential for hybridizing with wild species.

Reader eloquently argues that social history is important to understand agricultural systems and sustainability. His engaging account of the potato's journey, from the Andes to Europe and beyond, starts and ends in local communities where the tuber is still central to daily life. Andean cultures cultivate potatoes in poor-quality soils at high altitudes, mainly because

T. MORRISON/SOUTH AMERICAN PICTURES

Mash production: potatoes are the world's most chemically dependent crop.

commonly shared vision of the future and to create new networks that enable a society to invest its science and technology resources more wisely, harness the beneficial effects of innovation, and ameliorate its risks. Yet foresight may provide only one way to create these benefits. As the book describes, the French government sponsors a small number of comprehensive foresight activities. By contrast, in the United States, many groups — from the independent but government-funded National Academy of Sciences to numerous non-profit organizations — offer visions of the future and build networks around them. Clearly, these approaches reflect different political and cultural contexts. But the different visions may provide different strengths and weaknesses, for instance, offering coherent actions versus resilience to surprise. The book helps frame questions about, for example, which approaches governments should choose, but it does not answer such questions.

To call this volume a handbook may be premature. The word connotes an easily consulted reference that provides quick answers to those engaged in an activity. As the editors note, technology foresight remains a diverse and experimental practice whose theoretical foundations are poorly understood and whose successes have not yet moved from the anecdotal to the empirically grounded. Much remains to be learned. Meanwhile, *The Handbook of Technology Foresight* provides an important survey of current knowledge that will help governments use foresight to navigate these tumultuous times. ■

Robert Lempert is director of the RAND Frederick S. Pardee Center for Longer Range Global Policy and the Future Human Condition, 1776 Main Street, Santa Monica, California 90401, USA.
e-mail: lempert@rand.org

Potatoes and poverty

Propitious Esculent: The Potato in World History
by John Reader
William Heinemann: 2008. 315 pp. £18.99

Propitious Esculent is not just a book about potatoes; it is also about poverty. The two are linked by history, and in this very readable account, anthropologist and journalist John Reader shows us how.

The cultivated potato, *Solanum tuberosum*, is one of around 1,500 species in the flowering plant genus *Solanum*, which also includes the tomato, aubergine and woody nightshade. There are some 190 species of wild potatoes, all found in the Andes — from these a single species has been domesticated and spread throughout the world. Those who only see potatoes in heaps at supermarkets may be surprised that the crop comes from a flowering plant and is one of South America's greatest contributions to the European diet. The United Nations Food and Agriculture Organization has declared 2008 the International Year of the Potato to raise awareness of its importance.

Botanically, the potato is a tuber, a swollen piece of underground stem where the plant stores starch. Wild potato plants and local 'primitive' varieties have tubers, but they are often small and oddly shaped, bearing little physical resemblance to those from cultivated varieties. Reader cites recent research on the taxonomy and domestication of these varied and complicated plants. Disappointingly, he does not incorporate recent work on the genetics of the potato and its relatives.

Nor does the book sufficiently discuss the genetic modification of potatoes for control of disease, an important issue for food poverty. Potatoes are one of the most expensive food crops in terms of pest and disease control. Reader cites the potato as the "world's most chemically dependent crop — with the global cost of fungicides standing at [US]$2 billion per year". This astounding figure comes almost as a footnote at the end of a long chapter about the discovery of *Phytophthora infestans*, the agent of potato blight. The fungus, now associated with the devastation of world potato crops, was first discovered in grapevines. Reader's account of the disease, its discovery and its action is riveting, but potatoes are almost incidental to his story. Today, we easily see the connection between symptom and disease and can then search for causative agents: this was not so in the days when people thought microorganisms arose from spontaneous generation.

Reader does detail the development of blight-resistant potato varieties through the plant-breeding work of Redcliffe N. Salaman at the University of Cambridge, UK, in the early twentieth century. But he does not discuss more modern and controversial approaches. Disease control is being developed at the International Potato Center in Peru through 'true potato seed' potatoes — the crop is replanted using the seed from the original potato plant rather than vegetatively propagated from small pieces of tuber. These varieties have great promise for improving the gene pool for disease resistance, especially in the Andes, where genetically engineered potatoes cannot be used because of their potential for hybridizing with wild species.

Reader eloquently argues that social history is important to understand agricultural systems and sustainability. His engaging account of the potato's journey, from the Andes to Europe and beyond, starts and ends in local communities where the tuber is still central to daily life. Andean cultures cultivate potatoes in poor-quality soils at high altitudes, mainly because

T. MORRISON/SOUTH AMERICAN PICTURES

Mash production: potatoes are the world's most chemically dependent crop.

they have been excluded from richer agricultural land by large landowners and commercial elites. Social factors also influenced the Irish potato famine of the mid-nineteenth century. Potatoes can feed a family well from a very small plot of land, improving offspring survival and thus driving population growth. Pushed onto marginal land by large landowners, Irish peasants nevertheless thrived by growing potatoes; they were desperately poor, but not starving. When the potato blight hit Ireland, the resultant starvation killed more than a million Irish people and led to the emigration of millions more.

In his account of the Irish famine, Reader offers the central message of the book. Eliminating extreme hunger and poverty is one of the United Nations' Millennium Development Goals, but the history of the potato shows us that truly eliminating poverty means much more than ensuring the security of food supplies and avoiding hunger; social equity is equally, if not more, important. Science on its own is no panacea for solving social ills. ■

Sandra Knapp is a plant taxonomist in the Department of Botany, Natural History Museum, Cromwell Road, London SW7 5BD, UK.
e-mail: s.knapp@nhm.ac.uk

Building from the environment

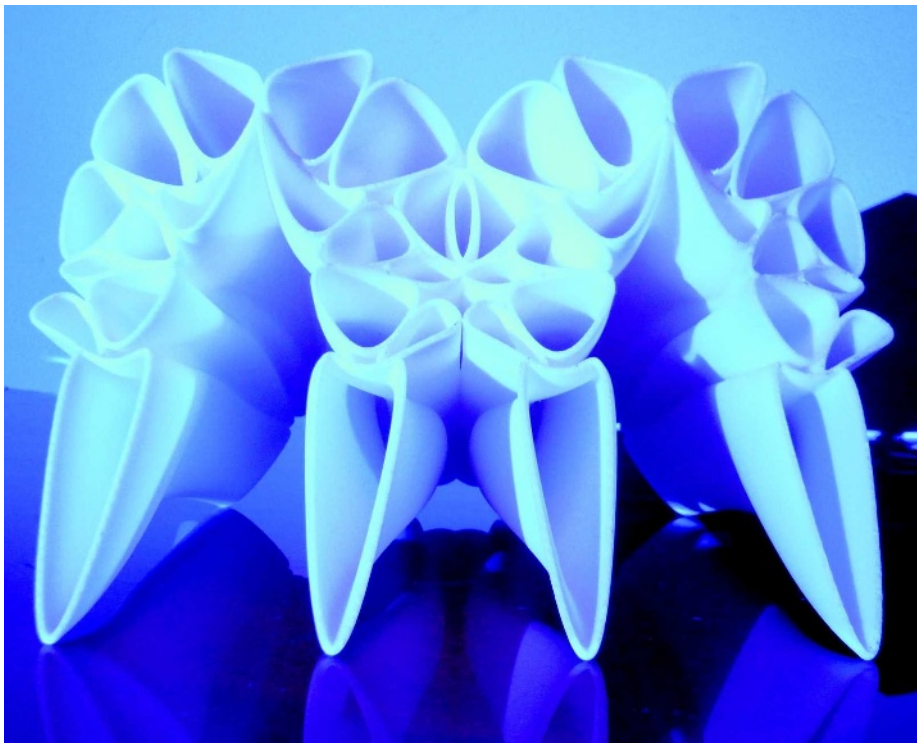
Abundant Australia
The Australian Pavilion, 11th International Architecture Exhibition, Venice
14 September until 23 November 2008

Moving beyond the creation of iconic buildings, architects are looking to natural forms for inspiration by appropriating biological patterns, structures and mechanisms to engage with the landscape. At the 11th International Architecture Exhibition at this year's Venice Biennale, as part of *Out There: Architecture Beyond Building*, several countries will focus on science in their installations. Japan picks the theme of extreme landscapes for its pavilion, Denmark highlights ecology, Egypt chooses geometry and Spain looks to a future of 'paperless' design and construction.

The Australian pavilion displays 300 model exhibits from 180 architecture and design practices, and highlights the inspiration that can be found around us with designs inspired by the flora, fauna and environment of that whole continent.

The microscopic structures of scales on moth and butterfly wings inspired the dual-layer, glazed facade on the Australian Museum's new Collections and Research Building, which will open in Sydney later this year. The cavity design of the panels will "insulate the building against extremes of temperature and humidity, and reduce traffic noise", says architectural practice Johnson Pilton Walker. Angled, dichroic glass within the panels seems to change colour as people walk past.

The Sydney-based French architect Frank Minnaërt asks how the functionality of biological membranes might be transposed into architecture, looking particularly at how organisms achieve maximum efficiency from minimal adaptation. His conceptual model *Patternity* comprises interacting units through which selectively permeable thermal membranes, yet to be developed, might stabilize building temperatures throughout the year.



Frank Minnaërt's model *Patternity* highlights the potential of membranes to cool new buildings.

Some designers are inspired by natural processes, such as fire and erosion, that alter the Australian environment. "We must conceive of architecture in a different way from the Western tradition based on a concept of heroic domination of space," says John Nichols of Woodhead architects, who incorporated scorched timber into the newly opened Pinnacles Desert Discovery Centre in Western Australia.

> **"We must conceive of architecture in a different way from the Western tradition based on a concept of heroic domination of space."**

Sunglasses for the Building, exhibited by spaceagency, is a light-hearted suggestion for providing shade for a beachside residential development under construction in Western Australia. "The screen detail was derived from images of nearby eroded limestone formations, and contributes to a unique sense of place," says practice director Michael Patroni.

These individual pavilions at the exhibition show that architectural responses to local environments can provide potent symbolism and functional solutions. As the exhibition's director Aaron Betsky states, "In a concrete sense, architecture is what allows us to be at home in the world." ■

Colin Martin is a writer based in London.
e-mail: cmpubrel@aol.com

they have been excluded from richer agricultural land by large landowners and commercial elites. Social factors also influenced the Irish potato famine of the mid-nineteenth century. Potatoes can feed a family well from a very small plot of land, improving offspring survival and thus driving population growth. Pushed onto marginal land by large landowners, Irish peasants nevertheless thrived by growing potatoes; they were desperately poor, but not starving. When the potato blight hit Ireland, the resultant starvation killed more than a million Irish people and led to the emigration of millions more.

In his account of the Irish famine, Reader offers the central message of the book. Eliminating extreme hunger and poverty is one of the United Nations' Millennium Development Goals, but the history of the potato shows us that truly eliminating poverty means much more than ensuring the security of food supplies and avoiding hunger; social equity is equally, if not more, important. Science on its own is no panacea for solving social ills. ■

Sandra Knapp is a plant taxonomist in the Department of Botany, Natural History Museum, Cromwell Road, London SW7 5BD, UK.
e-mail: s.knapp@nhm.ac.uk

Building from the environment

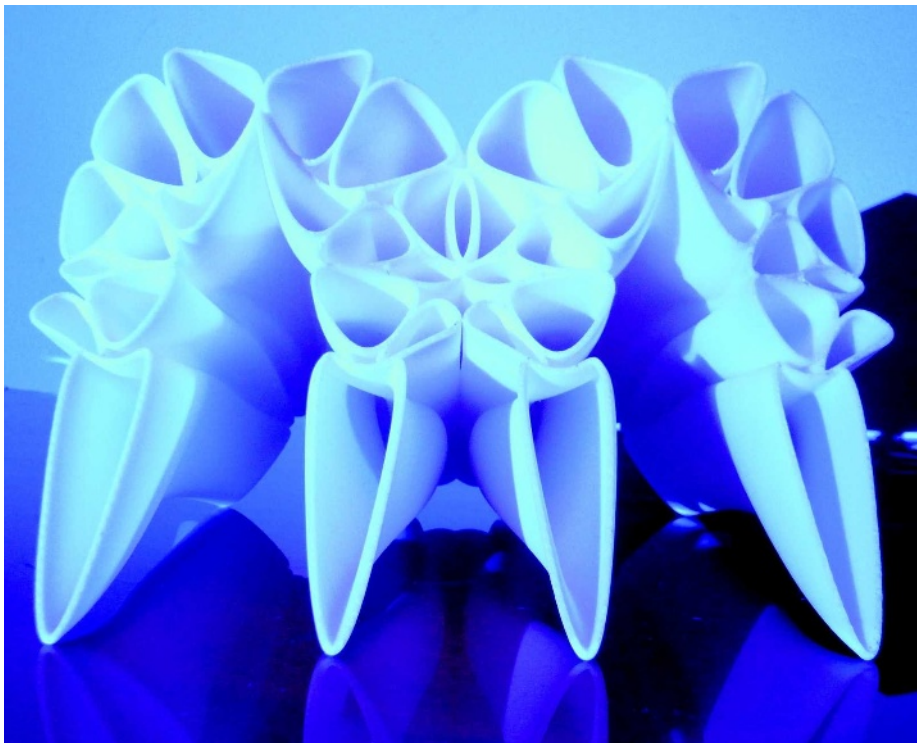
Abundant Australia
The Australian Pavilion, 11th International Architecture Exhibition, Venice
14 September until 23 November 2008

Moving beyond the creation of iconic buildings, architects are looking to natural forms for inspiration by appropriating biological patterns, structures and mechanisms to engage with the landscape. At the 11th International Architecture Exhibition at this year's Venice Biennale, as part of *Out There: Architecture Beyond Building*, several countries will focus on science in their installations. Japan picks the theme of extreme landscapes for its pavilion, Denmark highlights ecology, Egypt chooses geometry and Spain looks to a future of 'paperless' design and construction.

The Australian pavilion displays 300 model exhibits from 180 architecture and design practices, and highlights the inspiration that can be found around us with designs inspired by the flora, fauna and environment of that whole continent.

The microscopic structures of scales on moth and butterfly wings inspired the dual-layer, glazed facade on the Australian Museum's new Collections and Research Building, which will open in Sydney later this year. The cavity design of the panels will "insulate the building against extremes of temperature and humidity, and reduce traffic noise", says architectural practice Johnson Pilton Walker. Angled, dichroic glass within the panels seems to change colour as people walk past.

The Sydney-based French architect Frank Minnaërt asks how the functionality of biological membranes might be transposed into architecture, looking particularly at how organisms achieve maximum efficiency from minimal adaptation. His conceptual model *Patternity* comprises interacting units through which selectively permeable thermal membranes, yet to be developed, might stabilize building temperatures throughout the year.



Frank Minnaërt's model *Patternity* highlights the potential of membranes to cool new buildings.

Some designers are inspired by natural processes, such as fire and erosion, that alter the Australian environment. "We must conceive of architecture in a different way from the Western tradition based on a concept of heroic domination of space," says John Nichols of Woodhead architects, who incorporated scorched timber into the newly opened Pinnacles Desert Discovery Centre in Western Australia.

> **"We must conceive of architecture in a different way from the Western tradition based on a concept of heroic domination of space."**

Sunglasses for the Building, exhibited by spaceagency, is a light-hearted suggestion for providing shade for a beachside residential development under construction in Western Australia. "The screen detail was derived from images of nearby eroded limestone formations, and contributes to a unique sense of place," says practice director Michael Patroni.

These individual pavilions at the exhibition show that architectural responses to local environments can provide potent symbolism and functional solutions. As the exhibition's director Aaron Betsky states, "In a concrete sense, architecture is what allows us to be at home in the world." ■

Colin Martin is a writer based in London.
e-mail: cmpubrel@aol.com

Hidden treasures: the moulage museum in Zurich

Medical students still learn about skin diseases from hundreds of wax models that also record early cancer research and the ravages of syphilis, reports **Alison Abbott**.

The last moulageuse produced her final work in 1963. Elsbeth Stoiber, creator of wax models at the University Hospital in Zurich, fought against a waning interest in her art when colour photography became cheap in the 1950s. Thanks to her efforts, a fine collection of some 1,800 models has survived — now housed in a new light-filled museum that seems part sanctuary, part horror-movie props room.

Using wax to represent the manifestations of disease in three dimensions became popular in the late nineteenth century. Practitioners of this art, known as medical moulage, guarded their methods closely. Many took their secrets to the grave.

The general technique involved making a plaster cast directly on a patient's diseased skin, filling the cast with waxes and resins, adding different coloured waxes to mimic scabs and glass bubbles to look like blisters, and finally inserting hairs individually. The result preserved for posterity, and in sublime detail, the particular stages of a disease — a godsend for medical teaching. But the procedure was arduous and required great skill, so only rich clinics could produce models in large numbers.

Stoiber relinquished her methods reluctantly, first at a medical conference in 1979, where she outlined the main steps. In 1998, accepting with a heavy heart that she really was the end of the line, she inducted the curator of the Zurich moulage collection in her recipes, art and science.

The Zurich recipes have turned out to be the most durable. Other important collections, such as those in Paris and Tokyo, have not retained their true colours. And the Second World War also finished off or depleted many other collections, either through bombing, as in Dresden, or because of poor storage. At Guy's Hospital in London, for example, the world's oldest collection was bunkered during the war in damp cellars, and a fungus infested many pieces.

The moulage collection at the University of Zurich was started in 1917, shortly after Bruno Bloch founded a dermatology clinic there. After decades of doldrums, interest was again revived in the 1980s. At first, historians had a hard time making sense of the collection. The models were labelled only with patients' names and dates of birth, the relevant medical records having long since been destroyed. Fortunately, careful library research united many objects with their histories: in the moulage era, great value was placed on publishing long case studies.

MOULAGENMUSEUM, ZURICH

Wax casts immortalized skin diseases such as tuberous sclerosis in lifelike detail.

Further investigation revealed that the Zurich moulages had been made for medical research as well as for the more customary teaching function. Many pieces record rare side-effects of drugs, the consequences of emerging surgical techniques, and results of basic experiments.

Moulages showing symptoms of venereal diseases, such as syphilis and gonorrhoea, served several purposes. Models were made of the facial deformities of women infected with syphilis, who were locked into wards during treatment. These moulages were often used in alarming health propaganda to discourage promiscuity.

The incarcerated women also took part in research. A 17-year-old girl, for example, is immortalized in a moulage that shows skin reactions on her hands after deliberate infection with a foot fungus — an early demonstration of late-type sensitization of the immune system.

The Zurich collection records the first observations that X-rays, brought into medical use in 1897, can cause cancer as well as cure it. Medical expectations for X-rays were initially very high. One moulage shows a skin area scarred with psoriasis; a second model of the same area made some years after X-ray therapy shows the skin now covered with an oozing carcinoma. Novel surgical therapies, such as varicose vein removal and plastic surgery, are documented through to their long-term, often unhappy, consequences.

Other exhibits are a snapshot of early basic research into cancer, recording, for example, the long-term effects of painting tar onto the skin of mice, and exposing rabbits' ears to X-rays. Bloch also had moulages made of the experiments he did on himself, and proved that eczema can sometimes result from an allergic reaction to external agents. Particularly shocking are the powerfully realistic models of diseases, such as tuberculosis, that were prevalent before the advent of antibiotics.

Today, Zurich medical students still enjoy practising their red-spot recognition skills in the museum. They find the moulages more intuitive than pictures or images on computer screens for assessing subtle differences in the surface manifestation of diseases. The public may take a more voyeuristic look on Wednesday and Saturday afternoons. ■

Alison Abbott is *Nature*'s senior European correspondent.

See www.moulagen.ch for details.

For more Hidden treasures see www.nature.com/nature/focus/hiddentreasures

ESSAY

DAVID PARKINS

Paris 1951: The birth of CERN

François de Rose chaired the meeting that founded Europe's premier facility for experimental nuclear and particle research. Here he relives the five days of drama that changed the world of physics.

As a young French diplomat taking my first steps in international affairs, I had the privilege of representing my country for several years at a United Nations commission in the late 1940s. The United States, under the leadership of the financier and presidential adviser Bernard Baruch and the physicist Robert Oppenheimer, wanted the United Nations to be given oversight of all the world's nuclear weapons and nuclear power — the so-called Baruch plan. The plan failed, but as France was a keen supporter, it gave me the opportunity to work with Oppenheimer. We met frequently to discuss tactics and strategy and soon became friends.

One day, Oppenheimer told me of a problem that was very much on his mind. Most of America's best physicists, he said, had like him been trained, or had worked, in Europe's pre-war laboratories. He believed that Europe's shaken nations did not have the resources to rebuild their basic physics infrastructure. He felt they would no longer be able to remain scientific leaders unless they pooled their money and talent. Oppenheimer also believed that it would be "basically unhealthy" if Europe's physicists had to go to the United States or the Soviet Union to conduct their research.

> **"It would be basically unhealthy if Europe's physicists had to go to the United States or the Soviet Union to conduct their research."**

The solution, Oppenheimer felt, was to find a way to enable Europe's physicists to collaborate. When the United Nations commission ended, I returned to France, and raised the idea with our foreign minister Robert Schuman, one of the founders of the European Community. Schuman liked it and allowed me, together with Francis Perrin, then head of France's atomic energy commission, to seek the support of colleagues in other European capitals. Slowly the idea that would later become CERN began to take shape.

We had a mixed reception. There was a good deal of support, but some governments and scientists saw the project as too costly at a time when Europe's citizens were being asked to tighten their belts. Others feared it would take money away from individual national labs — which might, in turn, affect the project, because successful international cooperation needed national labs to be well resourced.

Still, by 1950 the project had gained considerable momentum and the American physicist Isidor Rabi had presented the idea to the member states of the United Nations Educational, Scientific and Cultural Organization (UNESCO) at an earlier meeting in

Florence, Italy. A date was then set for a follow-up meeting at UNESCO headquarters in Paris on 17 December 1951, where the idea would be debated and more details discussed.

View from the chair

I was asked to chair what would be perhaps the most important meeting in the history of CERN. It was attended by a who's who of twentieth-century physics. G. P. Thomson represented the United Kingdom, Francis Perrin spoke for France, Werner Heisenberg for Gemany and Jakob Nielsen and Niels Bohr represented Denmark. In all, 21 countries sent delegations, as did four international organizations, including the Council of Europe and the then International Council of Scientific Unions (now the International Council for Science). UNESCO was represented by the physicist Pierre Auger.

Delegates had many questions: did Europe really need a new and permanent experimental research facility, or would it be better if scientists collaborated in existing European labs? How much money would such a facility need? Which governments were prepared to pay, and how much would they pledge? Earlier disagreements soon became public as Germany and the United Kingdom, two nations whose support was critically needed, spoke out about their scepticism.

Auger opened by publicly thanking the United States for suggesting the idea to UNESCO. Next, Thomson rose to speak, and as the official report of the meeting records, he got straight to the point: "Britain has, since the war, spent a large sum of money on nuclear physics and especially on large machines. In the present state of financial stringency, further large expenditure by Britain on nuclear physics would not be justified. It must be remembered that there are other expensive branches of science which have a claim on our finances."

Thomson instead favoured the idea that Europe's physicists should collaborate using existing facilities. This would have the advantage that physicists could begin work immediately and not have to wait many years for a new facility to be completed. As a sign of the seriousness of his proposal, Thomson offered the use of a 400-MeV cyclotron at Liverpool University, which was nearing completion.

"The greatness of an institution is not to be measured only, or even mainly, by the size of its budget," he concluded. "Men are more important than machines." Later, Steva Dedijer, the delegate from Yugoslavia, countered: "Europe is supreme in knowing how to develop man. But men can't work without machines. And they will go where there are machines."

For France, Perrin said that the lack of more powerful equipment in the physics of fundamental particles would have the effect of "prejudicing European states and the aspects of civilization that they represent". He reminded the meeting that Europe's scientists would move to America if they couldn't find good facilities at home; and he said that building a machine comparable to those being constructed in the United States would be "far beyond the means of any single European state". Perrin advised that even if the United Kingdom's offer were to be accepted, work on the new laboratory should not be delayed.

The record of the meeting shows that influential backing for Thomson's view came from Heisenberg. "Our country is in an extremely difficult economic position and I am not entitled at the present time to commit our government to any expense in this connection," he said. He too emphasized that it was important that any scheme should produce results quickly and at minimum cost. "One should not just try to copy one of the big American machines." Nielsen, for Denmark, agreed that young researchers from Europe's scientifically less-developed countries were keen to begin work immediately using whatever experimental facilities were available.

> **"It was obvious that there were many diplomats among scientists."**

Yet, as the meeting progressed, it became clear that more delegates were in favour of building a new machine than against, and concrete offers of support started to come in. By the end, France, Switzerland, Italy, Belgium and Yugoslavia had collectively pledged $151,000 towards a feasibility study and Denmark said it would very probably join them. Denmark also proposed Copenhagen as a possible site for the new laboratory, with Belgium and Italy suggesting Geneva.

CERN takes shape

Two months later, 11 European governments agreed to establish a provisional governing council and the CERN acronym (Conseil Européen pour la Recherche Nucléaire) was born. Thanks to the generosity and farsightedness of Switzerland, Geneva was chosen as the site of the laboratory in October 1952, and in July 1953 the CERN Convention was ratified by the 12 founding member states: Belgium, Denmark, France, the Federal Republic of Germany, Greece, Italy, the Netherlands, Norway, Sweden, Switzerland, the United Kingdom and Yugoslavia.

The first cyclotron — a 600-MeV device — came into operation in 1957. Two years later it was joined by the 28-GeV proton synchrotron, which was for a brief period the world's highest-energy particle accelerator.

Today, as CERN enters an exciting new phase, it is worth recalling the many paradoxes in the foundation of this great institution. For example, at the 1951 meeting, unusually for the time, the United Kingdom took an opposite position to America's known wishes. Also unusual was the fact that the United States felt more strongly than Britain the need to strengthen European science, a major component of European culture.

Although early proponents of the idea of CERN also included the influential French physicist Louis de Broglie, it is impossible to overstate how important it was for all the proponents of CERN to have the United States take the lead and present the idea to UNESCO — this made the proposal much harder for others in Europe to oppose. But American support for CERN may have come at a price for American physicists. In later years, US policy-makers have used the existence of CERN as a reason to refuse requests from the US scientific community for expensive high-energy machines in their own country.

Few of us present that December in 1951 thought that by the time the meeting closed there would be so many pledges to take the idea of CERN forward. We began the meeting voicing different points of view, yet holding a unified vision for greater scientific cooperation and lasting peace in our continent. That vision is the one that eventually prevailed. Had the meeting failed, had scientists and governments not been able to agree on a joint programme of action, the repercussions of failure would have been felt far beyond the universe of nuclear physics.

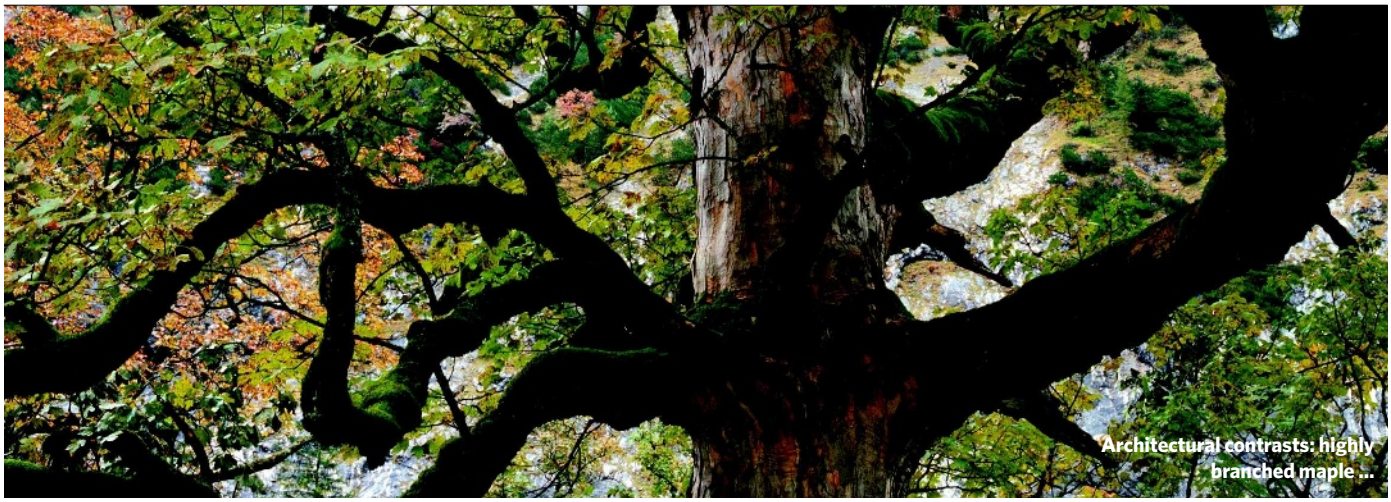
The meeting was a success, and this allowed me to close our deliberations with the remark that: "if it would be difficult to find scientists among diplomats, it was obvious that there were many diplomats among scientists." ■

François de Rose chaired the UNESCO meeting that was held in Paris from 17 to 21 December 1951. He was president of the council of CERN from 1959 to 1962 and was France's ambassador to NATO from 1970 to 1975. He is the author of *La France et la défense de l'Europe* (Seuil, 1976), translated as *European Security and France* (Macmillan, 1984).

See Editorial, page 137.

This is the first of a series: for more Meetings that Changed the World over the next five weeks see www.nature.com/nature/focus/meetings

NEWS & VIEWS



Architectural contrasts: highly branched maple ...

IMAGEBROKER.NET/PHOTOSHOT

PLANT BIOLOGY

Hormones branch out

Harry Klee

Evidence points to the existence of a hitherto uncharacterized type of hormone that controls different aspects of plant growth and interaction. The hunt for that hormone is heating up.

Plants can't move around and so have evolved elaborate biochemical communication systems to control growth in response to a changing environment. One of those systems involves the ability to make shoot branches. Branching habit defines the architecture of a plant, and elsewhere in this issue Gomez-Roldan *et al.*[1] (page 189) and Umehara *et al.*[2] (page 195) open a fresh avenue in the quest to find out precisely what those regulatory factors are.

Consider two tree species, maple and redwood, with greatly different architecture. A Californian redwood achieves great height because it is apically dominant; the apical shoot suppresses growth of subapical lateral shoots. By contrast, a maple is less apically dominant, has multiple growing shoot tips and becomes highly branched. Although architecture is largely determined by genetics, a plant must be able to modify its growth in response to the environment. If the dominant shoot is destroyed, for example, the plant responds by initiating growth of a subapical shoot bud.

Hormones are essential to the communication network that provides plants with growth plasticity. Two classes of hormone in particular, auxins and cytokinins, have long been known to influence apical dominance[3]. In recent years, genetic and biochemical evidence has implicated another class of hormone in branching control, one derived from carotenoids[4–6]. Plants that have mutations in genes encoding carotenoid-cleaving dioxygenases (CCDs) are highly branched, indicating that some substance normally suppresses the growth of lateral shoots[7,8]. Grafting and gene-expression studies indicate that the substance is produced principally in roots and is translocated to shoots, where it suppresses subapical shoot outgrowth. This substance, therefore, conforms to the classical definition of a hormone: it is produced in one tissue and translocated to another where it exerts a strong effect on growth.

Gomez-Roldan *et al.*[1] and Umehara *et al.*[2] report a considerable advance in identifying this new class of hormone. Between them, they have used common experimental plants — pea, *Arabidopsis* and rice — to show that levels of strigolactones, a group of terpenoid lactones thought to be derived from carotenoids, are significantly reduced in *ccd* branching mutants. The two studies are complementary in terms of their approaches and the test plants involved, and are consistent in their conclusions. Application of strigolactones to mutants restores normal branching. Crucial evidence comes from mutant plants that have a defect in the signalling pathway downstream of strigolactone. The defect is in a control component of the pathway, an F-box protein, which is postulated to transduce the hormone signal[9]. These mutants are not deficient in strigolactone synthesis and do not respond to application of strigolactone.

... and apically dominant redwood.

A. GEIGER/GETTY IMAGES

Strigolactones are compounds that stimulate seed germination in plants, such as *Striga*, that parasitize the roots of other plants[10]. They also act as signals for symbiotic interaction with the arbuscular mycorrhizal fungi[11] that colonize roots and facilitate the uptake of soil nutrients by plants. But the link with above-ground shoot branching was unexpected. Both groups[1,2] propose that strigolactones are themselves either hormones or their biosynthetic precursors. Although the teams' findings link the biosynthetic pathways of strigolactones with the elusive branching hormone, the details of

the pathway(s) have yet to be determined.

Strigolactones contain a large, four-ring backbone structure, probably derived from a carotenoid[12]. So far, three enzymes that might be involved in its synthesis have been identified by extensive mutation screenings of various plant species, but this is too few to synthesize such a complex structure. Nonetheless, applications of small amounts of strigolactone restore branching mutants to normal[1,2], indicating that if a strigolactone is not the actual hormone, it is very closely related to it.

Strigolactones are produced by the roots of many plants and the CCD genes are present in all higher plants. The involvement of strigolactones in mycorrhizal symbiosis suggests that they have a pivotal role in coordinating plant growth below as well as above ground. Mycorrhizal fungi promote root growth and, in turn, shoot growth. By extension, strigolactones could be the regulators that modulate appropriate shoot outgrowth. That parasitic plants in turn monitor such an influential root-produced compound is a marvellous example of co-evolution.

The identification of compounds that alter branching, mycorrhizal colonization and the germination of parasitic-plant seeds offers hope that customized chemicals can be designed to change these various responses. Species of *Striga* and *Orobanche* — another group of parasitic plants — cause massive crop losses in the developing world, especially in Africa. A cheap chemical that stimulates premature germination of these parasites would have immediate and widespread application. Similarly, chemicals that predictably alter plant architecture would be welcomed, particularly by the part of the horticultural industry that produces ornamental plants.

With these papers[1,2], we have moved closer to the identification of an entirely new class of plant hormone, and now have a biochemical handle on the control of several aspects of plant growth. Full characterization of the biologically active compounds that regulate branching should permit rapid progress in our understanding of the downstream signalling events, and of how this pathway interfaces with the auxin and cytokinin signalling pathways. ■

Harry Klee is in the Horticultural Sciences Department, University of Florida, Gainesville, Florida 32611-0690, USA.
e-mail: hjklee@ifas.ufl.edu

1. Gomez-Roldan, V. *et al. Nature* **455,** 189–194 (2008).
2. Umehara, M. *et al. Nature* **455,** 195–200 (2008).
3. McSteen, P. & Leyser, O. *Annu Rev. Plant Biol.* **56,** 353–374 (2005).
4. Beveridge, C. A., Ross, J. J. & Murfet, I. C. *Plant Physiol.* **104,** 953–959 (1994).
5. Sorefan, K. *et al. Genes Dev.* **17,** 1469–1474 (2003).
6. Booker, J. *et al. Curr. Biol.* **14,** 1232–1238 (2004).
7. Schwartz, S. H., Qin, X. & Loewen, M. C. *J. Biol. Chem.* **279,** 46940–46945 (2004).
8. Auldridge, M. E. *et al. Plant J.* **45,** 982–993 (2006).
9. Stirnberg, P., van de Sande, K. & Leyser, H. M. O. *Development* **129,** 1131–1141 (2002).
10. Cook, C. E. *et al. J. Am. Chem. Soc.* **94,** 6198–6199 (1972).
11. Akiyama, K. *et al. Nature* **435,** 824–827 (2005).
12. Matusova, R. *et al. Plant Physiol.* **139,** 920–934 (2005).

GAMMA-RAY BURSTS

Light on the distant Universe

Jonathan Grindlay

Observations of a long-lasting γ-ray burst, one that has the brightest optical counterpart yet discovered, challenge theoretical understanding of these bursts but may enhance their usefulness as cosmic probes.

On a clear night, from one of Earth's increasingly rare dark sites, one can see roughly 3,000 stars with the naked eye. All of these point sources of light are stars within our Milky Way galaxy, and most are closer than about 1,500 light years. It is only with the rare catastrophic end of a massive star's life, in a gargantuan explosion resulting from the collapse of the stellar core, that nature extends our visible reach with a supernova.

Possibly one in every thousand supernovae is not 'normal': as the core collapses past the state of a neutron star to a black hole, the spinning disk around the nascent black hole launches a powerful jet that 'drills' its way out of the overlying star[1] and produces an even more extreme blast: a long-duration γ-ray burst (GRB). These bursts typically last between 3 and 100 seconds, and are followed by fading afterglow emission at longer wavelengths (X-ray, optical, infrared and sometimes radio). On page 183 of this issue, Racusin *et al.*[2] report observations of the optically brightest GRB yet seen. The optical emission of this burst, dubbed GRB 080319B, is a hundred times brighter than the previous record holder.

GRB 080319B was detected by the Burst Alert Telescope (BAT) onboard NASA's Swift satellite on 19 March 2008. Only automated telescopes detected it, but it would have been visible to the naked eye for about 40 seconds — and thus whoever saw it would have witnessed the most distant astronomical object ever directly seen. Spectra of the optical afterglow measured its redshift as $z = 0.93$ (ref. 3), which corresponds to a light travel time of 7.4 billion years, placing GRB 080319B more than halfway back to the Big Bang and the origin of our Universe.

The only 'normal' supernova visible to the naked eye in the past 400 years, SN 1987A, was detected[4] on 24 February 1987. Its optical brightness was comparable to that of GRB 080319B, but it occurred a mere 163,000 light years away in our neighbouring satellite galaxy, the Large Magellanic Cloud. How could the similarly bright optical flash of GRB 080319B be in any way connected to the process of stellar death, given its approximately 5×10^4 times greater distance? The answer is 'beaming' — in which, instead of the isotropic, relatively slow emission from a normal supernova over days to months, a large fraction of the total energy of a GRB is collimated into a narrow and highly relativistic jet (that is, its bulk outflow velocity is very close to the speed of light).

Racusin and colleagues[2] show that the jet in GRB 080319B almost certainly has a two-component structure: a jet approximately 8° across surrounding a narrower (about 0.4°) central core of higher relativistic speed for which outflow velocities are within about five parts in ten million of the speed of light. For about 100 seconds, the collimated radiation beam observable from this jet was an intense beacon illuminating the intervening Universe. It came from a GRB that occurred around 3 billion years before the Sun and Earth formed.

X-ray, optical and radio observations of GRBs have shown that their afterglow emission is due to the collision of a beamed jet with the surrounding wind from the pre-supernova star and interstellar medium, and that beaming is directly indicated by the 'jet breaks' in the afterglow light curves[5]. Even more convincing evidence for the relativistic expansion of the jet was provided by the radio observations of another GRB — GRB 970508 — which showed[6] that its total energy was about ten times lower than inferred from a spherical explosion, implying a jet with an opening angle of about 30°. However, until the remarkably complete broadband spectral and temporal coverage of GRB 080319B, it had not been possible to directly constrain the radial structure of the jet.

Observations began before the BAT detection with optical imaging from wide-field telescopes that were already observing another burst, GRB 080319A, which was only 10° away from GRB 080319B and had gone off only 30 minutes before. This was a remarkable coincidence, given that the BAT observes only about two GRBs per week over the full sky. Ultimately, the afterglow from GRB 080319B was observed to fade by eight orders of magnitude in flux over six weeks by a worldwide suite of telescopes spanning 11 orders of magnitude in wavelength.

A prediction[2] of the high outflow velocities inferred for the central jet is the production of even more luminous, prompt GRB emissions of much higher-energy γ-rays. Such emissions would be easily detected by the recently launched Fermi Gamma-ray Space Telescope. But absorption of such high-energy γ-rays by the dense optical–ultraviolet photons produced by synchrotron emission in the same internal shock region could attenuate such emissions, despite the small angle scattering in the narrow jet.

The ultra-relativistic core of the jet in

the pathway(s) have yet to be determined.

Strigolactones contain a large, four-ring backbone structure, probably derived from a carotenoid[12]. So far, three enzymes that might be involved in its synthesis have been identified by extensive mutation screenings of various plant species, but this is too few to synthesize such a complex structure. Nonetheless, applications of small amounts of strigolactone restore branching mutants to normal[1,2], indicating that if a strigolactone is not the actual hormone, it is very closely related to it.

Strigolactones are produced by the roots of many plants and the CCD genes are present in all higher plants. The involvement of strigolactones in mycorrhizal symbiosis suggests that they have a pivotal role in coordinating plant growth below as well as above ground. Mycorrhizal fungi promote root growth and, in turn, shoot growth. By extension, strigolactones could be the regulators that modulate appropriate shoot outgrowth. That parasitic plants in turn monitor such an influential root-produced compound is a marvellous example of co-evolution.

The identification of compounds that alter branching, mycorrhizal colonization and the germination of parasitic-plant seeds offers hope that customized chemicals can be designed to change these various responses. Species of *Striga* and *Orobanche* — another group of parasitic plants — cause massive crop losses in the developing world, especially in Africa. A cheap chemical that stimulates premature germination of these parasites would have immediate and widespread application. Similarly, chemicals that predictably alter plant architecture would be welcomed, particularly by the part of the horticultural industry that produces ornamental plants.

With these papers[1,2], we have moved closer to the identification of an entirely new class of plant hormone, and now have a biochemical handle on the control of several aspects of plant growth. Full characterization of the biologically active compounds that regulate branching should permit rapid progress in our understanding of the downstream signalling events, and of how this pathway interfaces with the auxin and cytokinin signalling pathways. ■

Harry Klee is in the Horticultural Sciences Department, University of Florida, Gainesville, Florida 32611-0690, USA.
e-mail: hjklee@ifas.ufl.edu

1. Gomez-Roldan, V. *et al. Nature* **455,** 189–194 (2008).
2. Umehara, M. *et al. Nature* **455,** 195–200 (2008).
3. McSteen, P. & Leyser, O. *Annu Rev. Plant Biol.* **56,** 353–374 (2005).
4. Beveridge, C. A., Ross, J. J. & Murfet, I. C. *Plant Physiol.* **104,** 953–959 (1994).
5. Sorefan, K. *et al. Genes Dev.* **17,** 1469–1474 (2003).
6. Booker, J. *et al. Curr. Biol.* **14,** 1232–1238 (2004).
7. Schwartz, S. H., Qin, X. & Loewen, M. C. *J. Biol. Chem.* **279,** 46940–46945 (2004).
8. Auldridge, M. E. *et al. Plant J.* **45,** 982–993 (2006).
9. Stirnberg, P., van de Sande, K. & Leyser, H. M. O. *Development* **129,** 1131–1141 (2002).
10. Cook, C. E. *et al. J. Am. Chem. Soc.* **94,** 6198–6199 (1972).
11. Akiyama, K. *et al. Nature* **435,** 824–827 (2005).
12. Matusova, R. *et al. Plant Physiol.* **139,** 920–934 (2005).

GAMMA-RAY BURSTS

Light on the distant Universe

Jonathan Grindlay

Observations of a long-lasting γ-ray burst, one that has the brightest optical counterpart yet discovered, challenge theoretical understanding of these bursts but may enhance their usefulness as cosmic probes.

On a clear night, from one of Earth's increasingly rare dark sites, one can see roughly 3,000 stars with the naked eye. All of these point sources of light are stars within our Milky Way galaxy, and most are closer than about 1,500 light years. It is only with the rare catastrophic end of a massive star's life, in a gargantuan explosion resulting from the collapse of the stellar core, that nature extends our visible reach with a supernova.

Possibly one in every thousand supernovae is not 'normal': as the core collapses past the state of a neutron star to a black hole, the spinning disk around the nascent black hole launches a powerful jet that 'drills' its way out of the overlying star[1] and produces an even more extreme blast: a long-duration γ-ray burst (GRB). These bursts typically last between 3 and 100 seconds, and are followed by fading afterglow emission at longer wavelengths (X-ray, optical, infrared and sometimes radio). On page 183 of this issue, Racusin *et al.*[2] report observations of the optically brightest GRB yet seen. The optical emission of this burst, dubbed GRB 080319B, is a hundred times brighter than the previous record holder.

GRB 080319B was detected by the Burst Alert Telescope (BAT) onboard NASA's Swift satellite on 19 March 2008. Only automated telescopes detected it, but it would have been visible to the naked eye for about 40 seconds — and thus whoever saw it would have witnessed the most distant astronomical object ever directly seen. Spectra of the optical afterglow measured its redshift as $z = 0.93$ (ref. 3), which corresponds to a light travel time of 7.4 billion years, placing GRB 080319B more than halfway back to the Big Bang and the origin of our Universe.

The only 'normal' supernova visible to the naked eye in the past 400 years, SN 1987A, was detected[4] on 24 February 1987. Its optical brightness was comparable to that of GRB 080319B, but it occurred a mere 163,000 light years away in our neighbouring satellite galaxy, the Large Magellanic Cloud. How could the similarly bright optical flash of GRB 080319B be in any way connected to the process of stellar death, given its approximately 5×10^4 times greater distance? The answer is 'beaming' — in which, instead of the isotropic, relatively slow emission from a normal supernova over days to months, a large fraction of the total energy of a GRB is collimated into a narrow and highly relativistic jet (that is, its bulk outflow velocity is very close to the speed of light).

Racusin and colleagues[2] show that the jet in GRB 080319B almost certainly has a two-component structure: a jet approximately 8° across surrounding a narrower (about 0.4°) central core of higher relativistic speed for which outflow velocities are within about five parts in ten million of the speed of light. For about 100 seconds, the collimated radiation beam observable from this jet was an intense beacon illuminating the intervening Universe. It came from a GRB that occurred around 3 billion years before the Sun and Earth formed.

X-ray, optical and radio observations of GRBs have shown that their afterglow emission is due to the collision of a beamed jet with the surrounding wind from the pre-supernova star and interstellar medium, and that beaming is directly indicated by the 'jet breaks' in the afterglow light curves[5]. Even more convincing evidence for the relativistic expansion of the jet was provided by the radio observations of another GRB — GRB 970508 — which showed[6] that its total energy was about ten times lower than inferred from a spherical explosion, implying a jet with an opening angle of about 30°. However, until the remarkably complete broadband spectral and temporal coverage of GRB 080319B, it had not been possible to directly constrain the radial structure of the jet.

Observations began before the BAT detection with optical imaging from wide-field telescopes that were already observing another burst, GRB 080319A, which was only 10° away from GRB 080319B and had gone off only 30 minutes before. This was a remarkable coincidence, given that the BAT observes only about two GRBs per week over the full sky. Ultimately, the afterglow from GRB 080319B was observed to fade by eight orders of magnitude in flux over six weeks by a worldwide suite of telescopes spanning 11 orders of magnitude in wavelength.

A prediction[2] of the high outflow velocities inferred for the central jet is the production of even more luminous, prompt GRB emissions of much higher-energy γ-rays. Such emissions would be easily detected by the recently launched Fermi Gamma-ray Space Telescope. But absorption of such high-energy γ-rays by the dense optical–ultraviolet photons produced by synchrotron emission in the same internal shock region could attenuate such emissions, despite the small angle scattering in the narrow jet.

The ultra-relativistic core of the jet in

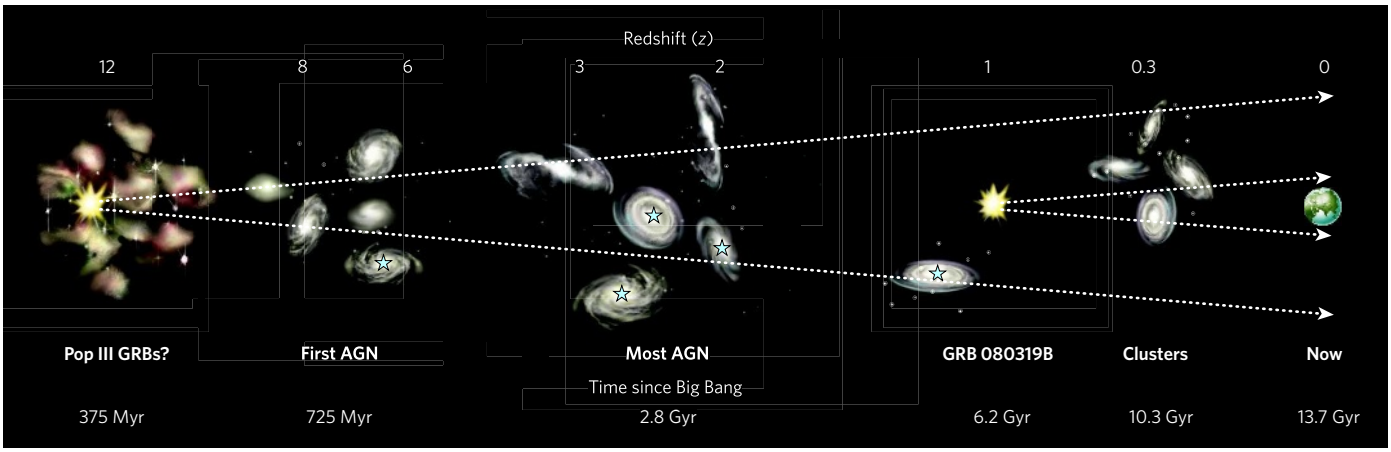


Figure 1 | Timeline of the Universe since the formation of the first stars. Dense clouds of gas collapse into the first (Pop III) massive stars and probably produce the first GRBs (ref. 7 and references therein). GRBs can then precede the formation of the first galaxies, which in turn precede that of active galactic nuclei (AGN) powered by supermassive black holes. Thus, GRBs could probe the first structures and galaxies to emerge after the 'dark ages' of the Universe. The narrow beaming of GRBs, best defined by GRB 080319B (not shown to scale), makes them the most luminous back-lights for mapping the far-distant visible Universe.

GRB 080319B challenges theories of jet formation and, more generally, models of the engine that drives GRBs. If these bursts contain radially structured jets, their usually inferred lower outflow velocities are explained by observations that are almost never exactly aligned with the narrow jet axis. Thus, the number of GRBs and their production rate, as well as the effects on their cosmic surroundings, may be a factor of 10–100 greater than previously thought.

This, in turn, bodes well for using the near-infrared spectra of GRBs to measure structure in the Universe at high redshift (Fig. 1). Because long GRBs are almost certainly produced by core collapse of massive stars to rapidly spinning black holes, and because such massive stars are expected to be the first stars formed in the Universe[7], the higher luminosity of jet cores, when aligned, should enable more-distant progenitors to be located by observations in the near infrared. A broader-band, wide-field, X-ray/γ-ray telescope is required that is designed to discover GRBs, and that is more sensitive and has higher spatial and spectral resolution than the BAT, to enable prompt near-infrared spectroscopy to be carried out. That would allow simultaneous GRB identification and redshift determination, as well as determination of the intervening cosmic structure. The Energetic X-ray Imaging Survey Telescope, EXIST, is a combined wide-field γ-ray imaging and optical–near-infrared imaging spectroscopy telescope[8] designed to discover and map the very first cosmic stellar black holes, and is under study by NASA for a proposed mission in the coming decade. ■

Jonathan Grindlay is at the Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts 02138, USA.
e-mail: josh@head.cfa.harvard.edu

1. MacFadyen, A. I., Woosley, S. E. & Heger, A. *Astrophys. J.* **550,** 410–425 (2001).
2. Racusin, J. L. *et al. Nature* **455,** 183–188 (2008).
3. Vreeswikj, P. M. *et al. GCN Circ.* **7444** (2008).
4. Kunkel, W. & Madore, B. *IAU Circ.* **4316** (1987).
5. Rhoads, J. E. *Astrophys. J.* **487,** L1–L4 (1997).
6. Frail, D. A. *et al. Astrophys. J.* **537,** 191–204 (2000).
7. Bromm, V. & Loeb, A. *AIP Conf. Proc.* **937,** 532–541 (2007).
8. Grindlay, J. E. *AIP Conf. Proc.* **836,** 631–641 (2006).

SCHIZOPHRENIA

Incriminating genomic evidence

James R. Lupski

The genetic factors that contribute to schizophrenia can vary, making it difficult to pinpoint which DNA changes are the main culprits. Large genome-wide studies provide the most reliable clues yet.

Schizophrenia is a chronic, debilitating illness with both neurological and psychiatric features, and it affects an estimated 1% of the world's population. Intense research into this disorder clearly points to the involvement of a significant genetic component, but genetic studies of schizophrenia have generally been disappointing as the data obtained often cannot be reproduced[1]. This lack of progress in understanding the genetic aspects of schizophrenia — which perhaps partially reflects challenges in diagnosis — is mainly due to the genetic heterogeneity among patients; many different genes might be involved in the disorder, but in a given family perhaps just one or a few of these genes mediate schizophrenia. Furthermore, previous studies have tended to be of limited statistical power and suboptimal design. Reporting in this issue, Stefansson *et al.*[2] and the International Schizophrenia Consortium[3] show that genome-wide studies of thousands of patients not only can confirm the association between previously identified genetic loci and the disease, but can also identify new loci.

Various genetic alterations can lead to disease. Examples include single nucleotide polymorphisms (SNPs) and gain or loss of large chunks of DNA known as copy-number variations (CNVs). The latter DNA rearrangements involve duplications and deletions that can result in many characteristics, including inherited neurological diseases and sporadic traits, and so are responsible for what are known as genomic disorders[4].

Stefansson *et al.*[2] (page 232) hypothesized that CNVs that confer a risk of disorders such as schizophrenia may be under negative-selection pressure because of the reduced fecundity of affected individuals. They therefore set out to identify *de novo* CNVs by searching for variations between the genomes of 9,878 sets of parents and offspring, none of whom was known to have schizophrenia. The authors identified 66 such CNVs, which they then tested for disease association in a sample of 1,433 patients with schizophrenia and related psychoses and 33,250 controls.

They found three genetic deletions that were nominally associated with schizophrenia and psychosis. These CNVs range in size from about 400 kilobases (kb) to 1.6 Mb, and are located on chromosomal regions 1q21.1, 15q11.2 and 15q13.3. A follow-up investigation with up to six other patient sample sets — in total consisting of 3,285 cases and 7,951 controls — revealed that, in the combined sample, all three deletions were significantly associated with schizophrenia and psychosis. The authors note that response rates to antipsychotic drugs

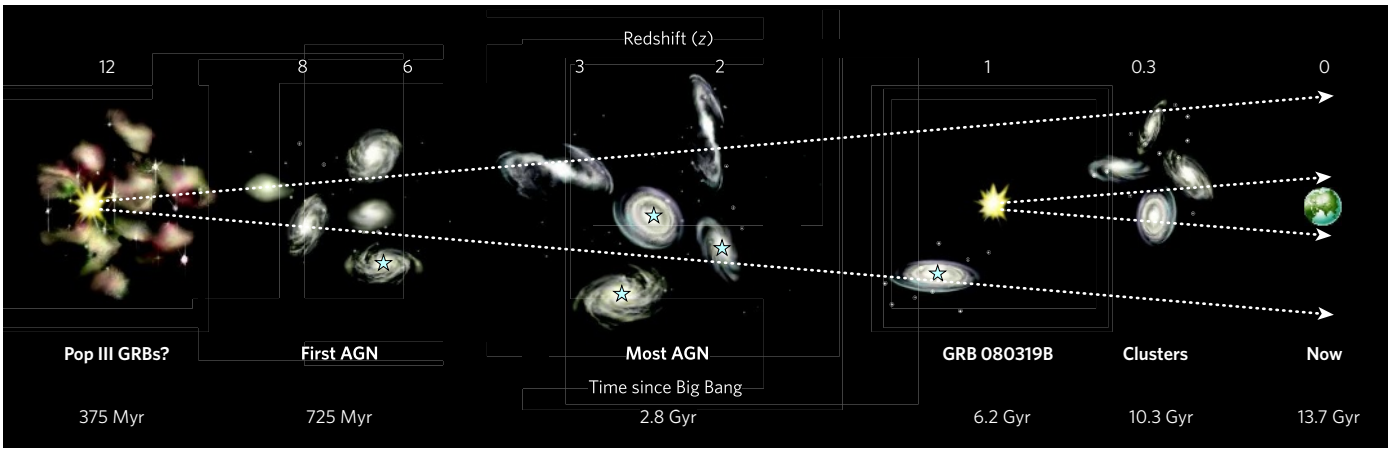


Figure 1 | Timeline of the Universe since the formation of the first stars. Dense clouds of gas collapse into the first (Pop III) massive stars and probably produce the first GRBs (ref. 7 and references therein). GRBs can then precede the formation of the first galaxies, which in turn precede that of active galactic nuclei (AGN) powered by supermassive black holes. Thus, GRBs could probe the first structures and galaxies to emerge after the 'dark ages' of the Universe. The narrow beaming of GRBs, best defined by GRB 080319B (not shown to scale), makes them the most luminous back-lights for mapping the far-distant visible Universe.

GRB 080319B challenges theories of jet formation and, more generally, models of the engine that drives GRBs. If these bursts contain radially structured jets, their usually inferred lower outflow velocities are explained by observations that are almost never exactly aligned with the narrow jet axis. Thus, the number of GRBs and their production rate, as well as the effects on their cosmic surroundings, may be a factor of 10–100 greater than previously thought.

This, in turn, bodes well for using the near-infrared spectra of GRBs to measure structure in the Universe at high redshift (Fig. 1). Because long GRBs are almost certainly produced by core collapse of massive stars to rapidly spinning black holes, and because such massive stars are expected to be the first stars formed in the Universe[7], the higher luminosity of jet cores, when aligned, should enable more-distant progenitors to be located by observations in the near infrared. A broader-band, wide-field, X-ray/γ-ray telescope is required that is designed to discover GRBs, and that is more sensitive and has higher spatial and spectral resolution than the BAT, to enable prompt near-infrared spectroscopy to be carried out. That would allow simultaneous GRB identification and redshift determination, as well as determination of the intervening cosmic structure. The Energetic X-ray Imaging Survey Telescope, EXIST, is a combined wide-field γ-ray imaging and optical–near-infrared imaging spectroscopy telescope[8] designed to discover and map the very first cosmic stellar black holes, and is under study by NASA for a proposed mission in the coming decade. ■

Jonathan Grindlay is at the Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts 02138, USA.
e-mail: josh@head.cfa.harvard.edu

1. MacFadyen, A. I., Woosley, S. E. & Heger, A. *Astrophys. J.* **550,** 410–425 (2001).
2. Racusin, J. L. *et al. Nature* **455,** 183–188 (2008).
3. Vreeswikj, P. M. *et al. GCN Circ.* **7444** (2008).
4. Kunkel, W. & Madore, B. *IAU Circ.* **4316** (1987).
5. Rhoads, J. E. *Astrophys. J.* **487,** L1–L4 (1997).
6. Frail, D. A. *et al. Astrophys. J.* **537,** 191–204 (2000).
7. Bromm, V. & Loeb, A. *AIP Conf. Proc.* **937,** 532–541 (2007).
8. Grindlay, J. E. *AIP Conf. Proc.* **836,** 631–641 (2006).

SCHIZOPHRENIA

Incriminating genomic evidence

James R. Lupski

The genetic factors that contribute to schizophrenia can vary, making it difficult to pinpoint which DNA changes are the main culprits. Large genome-wide studies provide the most reliable clues yet.

Schizophrenia is a chronic, debilitating illness with both neurological and psychiatric features, and it affects an estimated 1% of the world's population. Intense research into this disorder clearly points to the involvement of a significant genetic component, but genetic studies of schizophrenia have generally been disappointing as the data obtained often cannot be reproduced[1]. This lack of progress in understanding the genetic aspects of schizophrenia — which perhaps partially reflects challenges in diagnosis — is mainly due to the genetic heterogeneity among patients; many different genes might be involved in the disorder, but in a given family perhaps just one or a few of these genes mediate schizophrenia. Furthermore, previous studies have tended to be of limited statistical power and suboptimal design. Reporting in this issue, Stefansson *et al.*[2] and the International Schizophrenia Consortium[3] show that genome-wide studies of thousands of patients not only can confirm the association between previously identified genetic loci and the disease, but can also identify new loci.

Various genetic alterations can lead to disease. Examples include single nucleotide polymorphisms (SNPs) and gain or loss of large chunks of DNA known as copy-number variations (CNVs). The latter DNA rearrangements involve duplications and deletions that can result in many characteristics, including inherited neurological diseases and sporadic traits, and so are responsible for what are known as genomic disorders[4].

Stefansson *et al.*[2] (page 232) hypothesized that CNVs that confer a risk of disorders such as schizophrenia may be under negative-selection pressure because of the reduced fecundity of affected individuals. They therefore set out to identify *de novo* CNVs by searching for variations between the genomes of 9,878 sets of parents and offspring, none of whom was known to have schizophrenia. The authors identified 66 such CNVs, which they then tested for disease association in a sample of 1,433 patients with schizophrenia and related psychoses and 33,250 controls.

They found three genetic deletions that were nominally associated with schizophrenia and psychosis. These CNVs range in size from about 400 kilobases (kb) to 1.6 Mb, and are located on chromosomal regions 1q21.1, 15q11.2 and 15q13.3. A follow-up investigation with up to six other patient sample sets — in total consisting of 3,285 cases and 7,951 controls — revealed that, in the combined sample, all three deletions were significantly associated with schizophrenia and psychosis. The authors note that response rates to antipsychotic drugs

in patients with these deletions are comparable to those of the general population of patients with schizophrenia.

With the removal of psychoses other than schizophrenia — as defined by strict diagnostic criteria — the association with the 1q21.1 and 15q11.2 loci drops just below significance. But Stefansson and colleagues rightly argue that there is no reason for disease features associated with a particular CNV to be confined to the current boundaries of psychiatric-disorder classification. The investigators also examined their cohort for deletion in the 22q11.2 locus, which causes velo-cardio-facial syndrome (VCFS), a condition often associated with schizophrenia. This deletion was identified in eight of 3,838 cases (0.2%), but was absent in 39,299 controls.

The authors found no significant association between schizophrenia and the 54–166 SNPs that map within each of the three chromosomal deletions. It has long been established[5] that CNVs can result in the misinterpretation of marker genotypes such as SNPs, because every locus is presumed to be biallelic — a person usually inherits one gene allele (copy) from each parent, but a locus becomes triallelic with a duplication and monoallelic with a deletion. As a result, linkage (co-inheritance) analysis is distorted. Perhaps CNVs also reduce the informativeness of association studies when a biallelic SNP assay is used. Moreover, as the authors contend, the markers used for SNP analysis might lack the ability to tag them, and thus rare SNPs at these loci that are associated with schizophrenia could be missed.

In a related paper, the International Schizophrenia Consortium[3] (page 237) performed a genome-wide survey of rare CNVs in 3,391 patients with schizophrenia and in 3,181 ancestrally matched controls. The authors found that, compared with controls, the total number of CNVs that are observed in less than 1% of the combined sample, and that are more than 100 kb in length, is increased in patients with schizophrenia. Moreover, other CNVs — deletions of 12p11.23 and 16p12.1–p12.2 — were observed in four patients each. These results reproduce recent data[6,7] obtained through studying much smaller patient sample sets (150 in one study[6] and 152 in another[7]), which found that an increased number of CNVs was associated with schizophrenia. The advantage of the much larger disease-sample size in the consortium study is that the authors could search for specific CNVs associated with schizophrenia.

Among the schizophrenia-associated CNVs described in this paper[3], rarer, single-occurrence CNVs and those that affect genes are more prominent. Also, as anticipated, deletions of the VCFS-associated 22q11.2 locus were detected in 0.4% of the disease cases. What's more, large deletions in the 1q21.1 and 15q13.3 loci — the same genomic regions identified by Stefansson *et al.*[2] — corresponded to previously unknown associations with schizophrenia, which remained significant after genome-wide corrections for multiple testing. Overall, the results provide strong support for a model of schizophrenia that includes the effects of rare CNVs occurring both across the whole genome and at specific loci, such as 1q21.1 and 15q13.3.

It is exciting that both papers[2,3] identify not only an association of the known 22q11.2 deletion with schizophrenia, but also two previously unidentified deletions. These two loci are flanked by low copy repeats — DNA sequences that are highly susceptible to mutation[8]. Although relatively rare, the two newly discovered CNVs indicate that schizophrenia can be caused by specific genomic rearrangements, and so perhaps can be classified as a genomic disorder[4].

A previous study[9] of some 1,000 patients with mental retardation of unknown cause reported a recurrent deletion CNV in the 15q13.3 locus that was associated with a mental retardation and seizure syndrome in nine individuals, including one with autism. Intriguingly, among the eight controls carrying the 15q13.3 deletion in Stefansson and colleagues' study[2], there was also one autistic individual. So it seems that the same deletion CNV can increase the risk of a broad range of clinical mental disorders.

As is often the case, many questions remain. How frequently do these deletion CNVs occur *de novo* and how often are they inherited? How frequently do they cause schizophrenia? What are the dosage-sensitive genes located in the deletion CNVs within the 1q21.1 and 15q13.3 loci? Are they involved in specific neural networks or pathways? Can this genome-wide screening approach for identifying rare CNVs also detect other genomic regions and genes associated with psychiatric illness, to provide yet further insights into the biology of these disorders? Will correcting abnormalities in gene dosage by RNA interference or epigenetic methods provide therapeutic avenues worth exploring? Nonetheless, although cautious optimism is warranted in these early days, such studies[2,3,6,7], together with work on autism[10–13], point to one fact: study of even complex traits should include an evaluation of CNVs and other genomic structural changes. ■

James R. Lupski is in the Departments of Molecular and Human Genetics, and of Pediatrics, Baylor College of Medicine and Texas Children's Hospital, Houston, Texas 77030, USA.
e-mail: jlupski@bcm.tmc.edu

1. Burmeister, M., McInnis, M. G. & Zöllner, S. *Nature Rev. Genet.* **9,** 527–540 (2008).
2. Stefansson, H. *et al. Nature* **455,** 232–236 (2008).
3. The International Schizophrenia Consortium *Nature* **455,** 237–241 (2008).
4. Lupski, J. R. *Trends Genet.* **14,** 417–422 (1998).
5. Lupski, J. R. *et al. Cell* **66,** 219–232 (1991).
6. Walsh, T. *et al. Science* **320,** 539–543 (2008).
7. Xu, B. *et al. Nature Genet.* **40,** 880–885 (2008).
8. Lupski, J. R. *Nature Genet.* **39,** S43–S47 (2007).
9. Sharp, A. J. *et al. Nature Genet.* **40,** 322–328 (2008).
10. Sebat, J. *et al. Science* **316,** 445–449 (2007).
11. Weiss, L. A. *et al. N. Engl. J. Med.* **358,** 667–675 (2008).
12. Kumar, R. A. *et al. Hum. Mol. Genet.* **17,** 628–638 (2008).
13. Morrow, E. M. *et al. Science* **321,** 218–223 (2008).

50 YEARS AGO

Theoretically, cancer therapeutic agents should be able to differentiate between malignant and normal human cells and should be more toxic to the malignant cells ... Early experiments with guinea pig sera gave the unexpected finding that leukaemic lymphocytes were frequently more sensitive to inactivated (56° C, 30 min.) than to fresh sera ... [H]uman leukaemic lymphocytes were sensitive to toxic factors in normal rabbit sera and in inactivated guinea pig sera. The rabbit sera usually killed the leukaemic lymphocytes in a few hours by 'fixation' and killed normal lymphocytes in a few days by intranuclear vacuolization.

ALSO:

This Slimming Business. By Prof. John Yudkin — [An] excellent account of nutrition that should enable the non-scientific reader to appreciate the reasons for the condemnation of much of the published nonsense on dieting.

From *Nature* 13 September 1958.

100 YEARS AGO

The Influence of Alcohol and other Drugs on Fatigue. By Dr. W. H. R. Rivers — [T]he author details the results obtained in an experimental research on the influence of certain drugs—caffeine, alcohol, cocaine, strychnine, and tobacco—on muscular and mental fatigue ... Caffeine in moderate doses (about 0.3 gram of the citrate) increases the capacity for both muscular and mental work, the stimulating action persisting for some time, and not being followed by any depressant action. Excessive doses, however ... are followed by a depressant action so marked that the drug in such circumstances becomes an accelerator of fatigue ... Alcohol in small doses (5–10 c.c.) seems to produce little effect, in larger doses (20–40 c.c.) the action was variable; in a subject not used to alcohol, sweating, giddiness, and other symptoms often ensued ... The capacity for mental work on the whole seemed to be lowered.

From *Nature* 17 September 1908.

50 & 100 YEARS AGO

in patients with these deletions are comparable to those of the general population of patients with schizophrenia.

With the removal of psychoses other than schizophrenia — as defined by strict diagnostic criteria — the association with the 1q21.1 and 15q11.2 loci drops just below significance. But Stefansson and colleagues rightly argue that there is no reason for disease features associated with a particular CNV to be confined to the current boundaries of psychiatric-disorder classification. The investigators also examined their cohort for deletion in the 22q11.2 locus, which causes velo-cardio-facial syndrome (VCFS), a condition often associated with schizophrenia. This deletion was identified in eight of 3,838 cases (0.2%), but was absent in 39,299 controls.

The authors found no significant association between schizophrenia and the 54–166 SNPs that map within each of the three chromosomal deletions. It has long been established[5] that CNVs can result in the misinterpretation of marker genotypes such as SNPs, because every locus is presumed to be biallelic — a person usually inherits one gene allele (copy) from each parent, but a locus becomes triallelic with a duplication and monoallelic with a deletion. As a result, linkage (co-inheritance) analysis is distorted. Perhaps CNVs also reduce the informativeness of association studies when a biallelic SNP assay is used. Moreover, as the authors contend, the markers used for SNP analysis might lack the ability to tag them, and thus rare SNPs at these loci that are associated with schizophrenia could be missed.

In a related paper, the International Schizophrenia Consortium[3] (page 237) performed a genome-wide survey of rare CNVs in 3,391 patients with schizophrenia and in 3,181 ancestrally matched controls. The authors found that, compared with controls, the total number of CNVs that are observed in less than 1% of the combined sample, and that are more than 100 kb in length, is increased in patients with schizophrenia. Moreover, other CNVs — deletions of 12p11.23 and 16p12.1–p12.2 — were observed in four patients each. These results reproduce recent data[6,7] obtained through studying much smaller patient sample sets (150 in one study[6] and 152 in another[7]), which found that an increased number of CNVs was associated with schizophrenia. The advantage of the much larger disease-sample size in the consortium study is that the authors could search for specific CNVs associated with schizophrenia.

Among the schizophrenia-associated CNVs described in this paper[3], rarer, single-occurrence CNVs and those that affect genes are more prominent. Also, as anticipated, deletions of the VCFS-associated 22q11.2 locus were detected in 0.4% of the disease cases. What's more, large deletions in the 1q21.1 and 15q13.3 loci — the same genomic regions identified by Stefansson *et al.*[2] — corresponded to previously unknown associations with schizophrenia, which remained significant after genome-wide corrections for multiple testing. Overall, the results provide strong support for a model of schizophrenia that includes the effects of rare CNVs occurring both across the whole genome and at specific loci, such as 1q21.1 and 15q13.3.

It is exciting that both papers[2,3] identify not only an association of the known 22q11.2 deletion with schizophrenia, but also two previously unidentified deletions. These two loci are flanked by low copy repeats — DNA sequences that are highly susceptible to mutation[8]. Although relatively rare, the two newly discovered CNVs indicate that schizophrenia can be caused by specific genomic rearrangements, and so perhaps can be classified as a genomic disorder[4].

A previous study[9] of some 1,000 patients with mental retardation of unknown cause reported a recurrent deletion CNV in the 15q13.3 locus that was associated with a mental retardation and seizure syndrome in nine individuals, including one with autism. Intriguingly, among the eight controls carrying the 15q13.3 deletion in Stefansson and colleagues' study[2], there was also one autistic individual. So it seems that the same deletion CNV can increase the risk of a broad range of clinical mental disorders.

As is often the case, many questions remain. How frequently do these deletion CNVs occur *de novo* and how often are they inherited? How frequently do they cause schizophrenia? What are the dosage-sensitive genes located in the deletion CNVs within the 1q21.1 and 15q13.3 loci? Are they involved in specific neural networks or pathways? Can this genome-wide screening approach for identifying rare CNVs also detect other genomic regions and genes associated with psychiatric illness, to provide yet further insights into the biology of these disorders? Will correcting abnormalities in gene dosage by RNA interference or epigenetic methods provide therapeutic avenues worth exploring? Nonetheless, although cautious optimism is warranted in these early days, such studies[2,3,6,7], together with work on autism[10–13], point to one fact: study of even complex traits should include an evaluation of CNVs and other genomic structural changes. ■

James R. Lupski is in the Departments of Molecular and Human Genetics, and of Pediatrics, Baylor College of Medicine and Texas Children's Hospital, Houston, Texas 77030, USA.
e-mail: jlupski@bcm.tmc.edu

1. Burmeister, M., McInnis, M. G. & Zöllner, S. *Nature Rev. Genet.* **9,** 527–540 (2008).
2. Stefansson, H. *et al. Nature* **455,** 232–236 (2008).
3. The International Schizophrenia Consortium *Nature* **455,** 237–241 (2008).
4. Lupski, J. R. *Trends Genet.* **14,** 417–422 (1998).
5. Lupski, J. R. *et al. Cell* **66,** 219–232 (1991).
6. Walsh, T. *et al. Science* **320,** 539–543 (2008).
7. Xu, B. *et al. Nature Genet.* **40,** 880–885 (2008).
8. Lupski, J. R. *Nature Genet.* **39,** S43–S47 (2007).
9. Sharp, A. J. *et al. Nature Genet.* **40,** 322–328 (2008).
10. Sebat, J. *et al. Science* **316,** 445–449 (2007).
11. Weiss, L. A. *et al. N. Engl. J. Med.* **358,** 667–675 (2008).
12. Kumar, R. A. *et al. Hum. Mol. Genet.* **17,** 628–638 (2008).
13. Morrow, E. M. *et al. Science* **321,** 218–223 (2008).

50 YEARS AGO

Theoretically, cancer therapeutic agents should be able to differentiate between malignant and normal human cells and should be more toxic to the malignant cells ... Early experiments with guinea pig sera gave the unexpected finding that leukaemic lymphocytes were frequently more sensitive to inactivated (56° C, 30 min.) than to fresh sera ... [H]uman leukaemic lymphocytes were sensitive to toxic factors in normal rabbit sera and in inactivated guinea pig sera. The rabbit sera usually killed the leukaemic lymphocytes in a few hours by 'fixation' and killed normal lymphocytes in a few days by intranuclear vacuolization.

ALSO:

This Slimming Business. By Prof. John Yudkin — [An] excellent account of nutrition that should enable the non-scientific reader to appreciate the reasons for the condemnation of much of the published nonsense on dieting.

From *Nature* 13 September 1958.

100 YEARS AGO

The Influence of Alcohol and other Drugs on Fatigue. By Dr. W. H. R. Rivers — [T]he author details the results obtained in an experimental research on the influence of certain drugs—caffeine, alcohol, cocaine, strychnine, and tobacco—on muscular and mental fatigue ... Caffeine in moderate doses (about 0.3 gram of the citrate) increases the capacity for both muscular and mental work, the stimulating action persisting for some time, and not being followed by any depressant action. Excessive doses, however ... are followed by a depressant action so marked that the drug in such circumstances becomes an accelerator of fatigue ... Alcohol in small doses (5–10 c.c.) seems to produce little effect, in larger doses (20–40 c.c.) the action was variable; in a subject not used to alcohol, sweating, giddiness, and other symptoms often ensued ... The capacity for mental work on the whole seemed to be lowered.

From *Nature* 17 September 1908.

50 & 100 YEARS AGO

QUANTUM MECHANICS

Entangled families

Markus Aspelmeyer and Jens Eisert

Quantum entanglement comes in a rich variety of types and families if more than two particles are involved. Experiments with photons are opening up fresh ways to systematically study multi-particle entanglement.

In 1935, one of the founders of quantum mechanics, Erwin Schrödinger, coined the term 'entanglement' to describe two quantum systems that are intimately correlated — even more strongly than is classically possible. Although its impact on the foundations of quantum physics was envisaged from the beginning (entanglement is at the heart of his famous cat paradox), Schrödinger could not have anticipated that it would become a basic concept in quantum information processing. Today, we know that when three or more particles become entangled, many surprising features occur that are of relevance both for fundamental studies and for practical applications. Such multi-particle entanglement exists in increasingly complicated families of entangled states, but these have proved difficult to study. Writing in *Physical Review Letters*, Wieczorek *et al.*[1] demonstrate a simple experimental set-up capable of observing an entire family of states accessible to four entangled photons.

The concept of entanglement was originally applied to correlations between two distinct quantum systems, until the gedanken experiment by Greenberger, Horne and Zeilinger (GHZ)[2] established that entanglement between three or more constituents is conceptually different from the two-particle situation (Box 1). Since then, multi-particle entanglement has been recognized as a powerful resource in quantum information processing and communication, quantum imaging and metrology. For example, particular multipartite entangled states known as cluster or graph states[3] can function as a universal quantum computer (the computational power residing in the multipartite correlations). Other states can enhance the phase sensitivity to achieve the ultimate Heisenberg limit in spectroscopy[4]. So creating and controlling multipartite entanglement is high on the experimental physicists' 'to do' list.

Unlike entanglement between two particles, which has been achieved in numerous experiments using elementary quantum systems such as atoms, ions or photons, the experimental control of multi-particle entanglement is still in its infancy. So far, photons are the more versatile particles because they can be individually manipulated by simple optical devices such as phase plates and polarizers; additional interference between multi-photon states is used to create a hitherto unequalled variety of different multipartite entangled states. Examples include GHZ states for quantum non-locality tests, cluster states for quantum computing, or W-states and Dicke states that demonstrate the robustness of multipartite entanglement[5]. Still, these quantum states cover only a few of the types of multipartite entanglement that are possible. Ideally, one would like to have a single apparatus that 'tunes in' any wanted (multipartite) entangled state by simply turning a knob. Each experiment performed so far, however, was designed to optimally generate specific entangled states by using specialized arrangements of phase plates, leaving little room for creating and distinguishing whole families of states.

Wieczorek and colleagues' experimental scheme[1] promises to overcome the inflexibility of previous linear-optical experiments. In their set-up, a single waveplate is sufficient to tune the generated quantum states through a whole family of 4-photon entangled states — one of the nine different families of states into which four particles can be entangled[6]. In essence, they use the emission from a parametric downconversion process in which an ultraviolet photon is converted into two polarization-entangled photons with the same combined energy and momentum as the original photon. Double-pair emission occurs as a second-order event and creates a 4-photon entangled state. The photons pass through the tuning waveplate, which changes the relative amplitudes of the different 4-photon contributions, and are then made to interfere at polarizing beamsplitters. Each waveplate orientation corresponds to a well-defined 4-particle entangled state. The entanglement fidelities achieved for several of the states in these experiments are comparable to previous experiments designed to produce a single state rather than all members of a family.

The results[1] show that the generation of photonic entanglement provides more flexibility than previously expected. But there are still limitations that hinder the generalization of such systematic studies to a large number of particles. It seems almost a general rule that current experiments with full control over individual particles (photons or ions) are limited to small numbers. This is mainly because of inefficient detection (as in the case of photons) or errors in the individual operations (as in the case of ions). At present, the

Box 1 | Multi-particle entanglement at a glance

The essence of entangled states is that they cannot be prepared with local physical devices. Outcomes of local measurements can be more strongly correlated than in any classical system: intriguingly, there can be no local 'catalogue' that fully determines these outcomes.

This counterintuitive trait of quantum mechanics is exploited in numerous applications of quantum information processing, as well as in explorations of the foundations of the theory. A pure state of a multi-particle quantum system is called entangled if it is not the product of states of each subsystem or, for general quantum states, if it is not a mixture of product states.

Two particles can only be purely entangled in one specific way, but entanglement between three or more particles is much richer. Indeed, depending on exactly how two quantum states are viewed to be equivalent, an entire zoo of different entanglement classes emerges in the multipartite setting[14,15].

The prototypical multi-particle state of three spin-$^1/_2$ systems (qubits) is the Greenberger–Horne–Zeilinger (GHZ) state represented by the state vector $|0,0,0\rangle+|1,1,1\rangle$ (that is, all three particles involved are in the same spin state). GHZ states allow formulation of a direct incompatibility between quantum physics and objective local theories without invoking Bell's inequalities.

The GHZ is fundamentally different from a bipartite entangled state in that it cannot be reversibly prepared from bipartite entangled states. Other multipartite entangled states, for example the W-state $|0,0,1\rangle+|0,1,0\rangle+|1,0,0\rangle$, are less strongly correlated but demonstrate robust entanglement (that is, some entanglement is preserved if a particle is removed from the state).

To grasp the many different classes of multipartite states, 'theoretical laboratories' have proved useful; these capture essential features, but are easier to understand in their properties. For example, GHZ states are instances of graph states, a class of multipartite entangled states that can be interpreted using mathematical graph theory — the relationships between the particles (vertices in the graphs) are represented by edges connecting them in multidimensional space.

A subset of graph states, cluster states, are states that can be fitted to a regular cubic lattice. Cluster states are a 'universal resource' for quantum computing: by making appropriate measurements of individual spins, any quantum-computer circuit can be efficiently simulated. Other theoretical laboratories used in classifications include matrix-product states, or Gaussian bosonic and fermionic states. Indeed, multi-particle entanglement presumably also occurs naturally in complex quantum systems. **M.A. & J.E.**

BIOMOLECULAR ENGINEERING

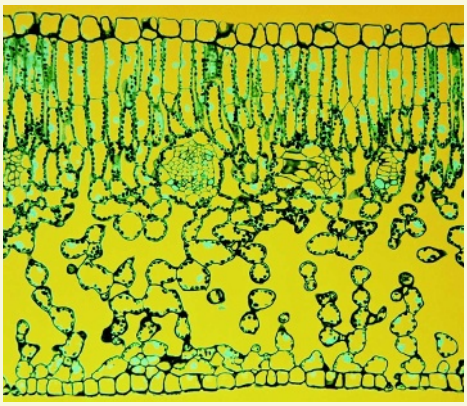
Negative success in tiny tree

When engineers were building beam engines in the early eighteenth century to pump out water-logged mines, they found that they couldn't pull water up more than about 9 metres (the height of water that can be supported by the drop in pressure between the atmosphere and a vacuum). Trees grow many times taller — more than 100 metres in the case of the tallest redwoods. Yet they supply their leaves with a constant flow of water. They achieve this feat by keeping the water high up in their trunks under pressures many atmospheres below that of a vacuum.

Elsewhere in this issue, Wheeler and Stroock report a duplication of this trick: they have created a tiny 'synthetic tree' through whose trunk water flows at pressures of around −10 atmospheres (T. D. Wheeler and A. D. Stroock *Nature* **455,** 208–212; 2008).

In trees, evaporation of water from leaf cells called spongy mesophyll pulls water up through hollow cells in the trunk (spongy mesophyll is the tissue in the lower half of this picture, a cross-section through a leaf). The strong, cohesive properties of water, responsible for its powerful surface tension, allow the water to exist at large negative pressures. But even the smallest bubble would explosively expand into the water, disrupting its flow in a process known as cavitation. The interface between the plant's water system and the air, formed by the spongy mesophyll, must allow water to pass, but not the gas molecules that would cause cavitation.

To create their tree, Wheeler and Stroock use a hydrogel, which mimics the mesophyll by holding water in molecular-scale pores, smaller than those of other porous solids. As their respective 'root' and 'leaf', the authors formed two networks of channels, 10 micrometres in diameter, in a sheet of poly(hydroxyethyl methacrylate), and connected them by a single channel, the 'trunk'. With the 'root' exposed to a source of water and the 'leaf' to a stream of damp air, water flows through the system powered solely by 'leaf' evaporation. The pressures developed in the trunk are some 15 times more negative than in any previously reported pumping system.



S. WALKER/GETTY

The device is shown in Figure 3a of the paper (page 210). It is just 5 centimetres long, and the flow is a little over 2 micrograms of water per second — but from such small acorns do mighty oaks grow. The synthetic tree can provide a test device for theories of tree physiology and, scaled-up, the technology could find uses in passive pumps or cooling devices — evaporation makes the 'leaf' a heat sink. Also, the large negative pressures developed might be used to drag water out of even quite dry soils, simultaneously filtering out impurities by passage through the 'root' hydrogel. This process, which the authors dub "reverse reverse osmosis", could form the basis of solar-powered mining of pure water in arid or contaminated environments.

Christopher Surridge

largest such entangled states contain eight ions[7] or six photons[8]. By contrast, experiments with entangled states that involve a large particle number (greater than 100) lack the possibility of controlling individual particles, which is necessary to exploit multipartite correlations. For example, nearest-neighbour interactions of atoms held in an optical lattice, produced by the interference of several laser beams, can create the entanglement of many thousands of atoms, but only as a result of global operations performed jointly on all atoms[9]. Manipulation and detection of single atoms within the lattice are not yet possible, because the lattice spacings are too narrow for any optical beam to interact with individual atoms.

Technological improvements, stimulated by the rapidly developing field of quantum information processing[10], will doubtless overcome these limitations in time. Brighter photon sources are being developed along with high-efficiency detectors with single-photon resolution. Continuous-variable approaches offer further possibilities beyond single-photon architectures. In addition, chip-based quantum circuits have been tested that contain silica-on-silicon waveguides[11]. Chip-based ion traps are also a promising route to engineering multipartite ion entanglement[7], and advances in single-atom manipulation within optical lattices[9] could provide controllable multipartite entanglement of thousands of particles.

The increased diversity of experimentally 'tamed' entangled particles brings theoretical challenges. The number of parameters needed to describe a state increases exponentially with the number of particles in the state, but this increased complexity also increases the potential utility of multipartite entangled states. For example, point-to-point quantum communication is more efficient when exploiting multipartite networks[12]. It is also possible that the computational power of multipartite states can be optimized or adapted to given architectures by making explicit use of specific entanglement properties such as intricate long-range correlations[13].

We are only beginning to understand and exploit the power of multipartite entanglement. Current developments may seem to be tiny steps, but they could soon add up to a (quantum) leap not only in information science but also in our fundamental understanding of macroscopic quantum systems. ■

Markus Aspelmeyer is at the Institute for Quantum Optics and Quantum Information, Austrian Academy of Sciences, 1090 Vienna, Austria. Jens Eisert is in the Department of Physics, University of Potsdam, 14469 Potsdam, Germany, and Imperial College London, UK.
e-mails: markus.aspelmeyer@quantum.at; jense@qipc.org

1. Wieczorek, W. *et al. Phys. Rev. Lett.* **101,** 010503 (2008).
2. Greenberger, D. M., Horne, M. A., Shimony, A. & Zeilinger, A. *Am. J. Phys.* **58,** 1131–1143 (1990).
3. Raussendorf, R. & Briegel, H. J. *Phys. Rev. Lett.* **86,** 5188–5191 (2001).
4. Leibfried, D. *et al. Science* **304,** 1476–1478 (2004).
5. Pan, J.-W., Chen, Z.-B., Zukowski, M., Weinfurter, H. & Zeilinger, A. preprint at http://arxiv.org/0805.2853 (2008).
6. Verstraete, F., Dehaene, J., De Moor, B. & Verschelde, H. *Phys. Rev. A* **65,** 052112 (2002).
7. Blatt, R. & Wineland, D. *Nature* **453,** 1008–1015 (2008).
8. Lu, C.-Y. *et al. Nature Phys.* **3,** 91–95 (2007).
9. Bloch, I. *Nature* **453,** 1016–1022 (2008).
10. Walmsley, I. A. *Science* **319,** 1211–1213 (2008).
11. Politi, A., Cryan, M. J., Rarity, J. G., Yu, S. & O'Brien, J. L. *Science* **320,** 646–649 (2008).
12. Acín, A., Cirac, J. I. & Lewenstein, M. *Nature Phys.* **3,** 256–259 (2007).
13. Gross, D. & Eisert, J. *Phys. Rev. Lett.* **98,** 220503 (2007).
14. Eisert, J. & Gross, D. preprint at http://arxiv.org/abs/quant-ph/0505149 (2005).
15. Plenio, M. B. & Virmani, S. *Quant. Inf. Comp.* **7,** 001–051 (2007).

Correction

The News & Views article "Materials science: A desirable wind up" (*Nature* **454,** 591–592; 2008), by Neil Mathur, considered work by D. Lebeugle *et al.* describing investigation of the multiferroic and magnetoelectric properties of single crystals of $BiFeO_3$ (*Phys. Rev. Lett.* **100,** 227602; 2008). Coverage of closely related work by V. Kiryukhin and colleagues (S. Lee *et al. Appl. Phys. Lett.* **92,** 192906; 2008), published just before that by Lebeugle *et al.*, was inadvertently omitted from the article.

BIOMOLECULAR ENGINEERING

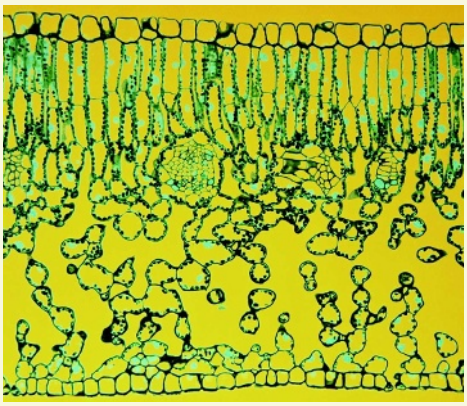
Negative success in tiny tree

When engineers were building beam engines in the early eighteenth century to pump out water-logged mines, they found that they couldn't pull water up more than about 9 metres (the height of water that can be supported by the drop in pressure between the atmosphere and a vacuum). Trees grow many times taller — more than 100 metres in the case of the tallest redwoods. Yet they supply their leaves with a constant flow of water. They achieve this feat by keeping the water high up in their trunks under pressures many atmospheres below that of a vacuum.

Elsewhere in this issue, Wheeler and Stroock report a duplication of this trick: they have created a tiny 'synthetic tree' through whose trunk water flows at pressures of around −10 atmospheres (T. D. Wheeler and A. D. Stroock *Nature* **455,** 208–212; 2008).

In trees, evaporation of water from leaf cells called spongy mesophyll pulls water up through hollow cells in the trunk (spongy mesophyll is the tissue in the lower half of this picture, a cross-section through a leaf). The strong, cohesive properties of water, responsible for its powerful surface tension, allow the water to exist at large negative pressures. But even the smallest bubble would explosively expand into the water, disrupting its flow in a process known as cavitation. The interface between the plant's water system and the air, formed by the spongy mesophyll, must allow water to pass, but not the gas molecules that would cause cavitation.

To create their tree, Wheeler and Stroock use a hydrogel, which mimics the mesophyll by holding water in molecular-scale pores, smaller than those of other porous solids. As their respective 'root' and 'leaf', the authors formed two networks of channels, 10 micrometres in diameter, in a sheet of poly(hydroxyethyl methacrylate), and connected them by a single channel, the 'trunk'. With the 'root' exposed to a source of water and the 'leaf' to a stream of damp air, water flows through the system powered solely by 'leaf' evaporation. The pressures developed in the trunk are some 15 times more negative than in any previously reported pumping system.



S. WALKER/GETTY

The device is shown in Figure 3a of the paper (page 210). It is just 5 centimetres long, and the flow is a little over 2 micrograms of water per second — but from such small acorns do mighty oaks grow. The synthetic tree can provide a test device for theories of tree physiology and, scaled-up, the technology could find uses in passive pumps or cooling devices — evaporation makes the 'leaf' a heat sink. Also, the large negative pressures developed might be used to drag water out of even quite dry soils, simultaneously filtering out impurities by passage through the 'root' hydrogel. This process, which the authors dub "reverse reverse osmosis", could form the basis of solar-powered mining of pure water in arid or contaminated environments.

Christopher Surridge

largest such entangled states contain eight ions[7] or six photons[8]. By contrast, experiments with entangled states that involve a large particle number (greater than 100) lack the possibility of controlling individual particles, which is necessary to exploit multipartite correlations. For example, nearest-neighbour interactions of atoms held in an optical lattice, produced by the interference of several laser beams, can create the entanglement of many thousands of atoms, but only as a result of global operations performed jointly on all atoms[9]. Manipulation and detection of single atoms within the lattice are not yet possible, because the lattice spacings are too narrow for any optical beam to interact with individual atoms.

Technological improvements, stimulated by the rapidly developing field of quantum information processing[10], will doubtless overcome these limitations in time. Brighter photon sources are being developed along with high-efficiency detectors with single-photon resolution. Continuous-variable approaches offer further possibilities beyond single-photon architectures. In addition, chip-based quantum circuits have been tested that contain silica-on-silicon waveguides[11]. Chip-based ion traps are also a promising route to engineering multipartite ion entanglement[7], and advances in single-atom manipulation within optical lattices[9] could provide controllable multipartite entanglement of thousands of particles.

The increased diversity of experimentally 'tamed' entangled particles brings theoretical challenges. The number of parameters needed to describe a state increases exponentially with the number of particles in the state, but this increased complexity also increases the potential utility of multipartite entangled states. For example, point-to-point quantum communication is more efficient when exploiting multipartite networks[12]. It is also possible that the computational power of multipartite states can be optimized or adapted to given architectures by making explicit use of specific entanglement properties such as intricate long-range correlations[13].

We are only beginning to understand and exploit the power of multipartite entanglement. Current developments may seem to be tiny steps, but they could soon add up to a (quantum) leap not only in information science but also in our fundamental understanding of macroscopic quantum systems. ■

Markus Aspelmeyer is at the Institute for Quantum Optics and Quantum Information, Austrian Academy of Sciences, 1090 Vienna, Austria. Jens Eisert is in the Department of Physics, University of Potsdam, 14469 Potsdam, Germany, and Imperial College London, UK.
e-mails: markus.aspelmeyer@quantum.at; jense@qipc.org

1. Wieczorek, W. *et al. Phys. Rev. Lett.* **101,** 010503 (2008).
2. Greenberger, D. M., Horne, M. A., Shimony, A. & Zeilinger, A. *Am. J. Phys.* **58,** 1131–1143 (1990).
3. Raussendorf, R. & Briegel, H. J. *Phys. Rev. Lett.* **86,** 5188–5191 (2001).
4. Leibfried, D. *et al. Science* **304,** 1476–1478 (2004).
5. Pan, J.-W., Chen, Z.-B., Zukowski, M., Weinfurter, H. & Zeilinger, A. preprint at http://arxiv.org/0805.2853 (2008).
6. Verstraete, F., Dehaene, J., De Moor, B. & Verschelde, H. *Phys. Rev. A* **65,** 052112 (2002).
7. Blatt, R. & Wineland, D. *Nature* **453,** 1008–1015 (2008).
8. Lu, C.-Y. *et al. Nature Phys.* **3,** 91–95 (2007).
9. Bloch, I. *Nature* **453,** 1016–1022 (2008).
10. Walmsley, I. A. *Science* **319,** 1211–1213 (2008).
11. Politi, A., Cryan, M. J., Rarity, J. G., Yu, S. & O'Brien, J. L. *Science* **320,** 646–649 (2008).
12. Acín, A., Cirac, J. I. & Lewenstein, M. *Nature Phys.* **3,** 256–259 (2007).
13. Gross, D. & Eisert, J. *Phys. Rev. Lett.* **98,** 220503 (2007).
14. Eisert, J. & Gross, D. preprint at http://arxiv.org/abs/quant-ph/0505149 (2005).
15. Plenio, M. B. & Virmani, S. *Quant. Inf. Comp.* **7,** 001–051 (2007).

Correction

The News & Views article "Materials science: A desirable wind up" (*Nature* **454,** 591–592; 2008), by Neil Mathur, considered work by D. Lebeugle *et al.* describing investigation of the multiferroic and magnetoelectric properties of single crystals of $BiFeO_3$ (*Phys. Rev. Lett.* **100,** 227602; 2008). Coverage of closely related work by V. Kiryukhin and colleagues (S. Lee *et al. Appl. Phys. Lett.* **92,** 192906; 2008), published just before that by Lebeugle *et al.*, was inadvertently omitted from the article.

OBITUARY

Neil Bartlett (1932–2008)

Founder of noble-gas chemistry.

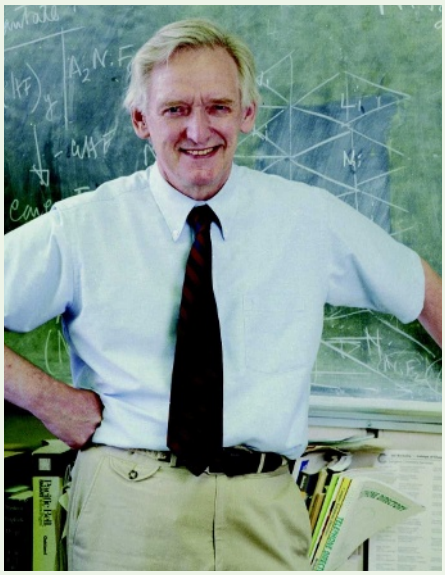
Neil Bartlett, who died on 5 August at the age of 75 from an aortic aneurysm, was one of the foremost chemists of the twentieth century. His discovery in the early 1960s of xenon fluorides, the first examples of a noble-gas compound, was a sensation. With one stroke of genius, he destroyed the long-standing dogma that the 'noble gases' (also previously known as the inert or rare gases) are unreactive.

Bartlett was born in 1932, in Newcastle upon Tyne, UK, and studied for both his undergraduate and postgraduate degrees at the University of Durham. Following a brief spell as a chemistry teacher, in 1958 he emigrated to Canada and took up a position as a lecturer at the University of British Columbia (UBC) in Vancouver. In 1964, he became a full professor at UBC, but in 1966 moved to a professorship at Princeton University and also joined the research staff at Bell Telephone Laboratories in Murray Hill, New Jersey. In 1969, he accepted a professorship at the University of California, Berkeley, and became a faculty senior scientist at the Lawrence Berkeley National Laboratory, positions he held until his retirements in 1993 and 1999, respectively.

Attempts to prepare compounds of the noble gases date back to the discovery of argon by William Ramsay in 1894. Among the ensuing notable events was Walther Kossel's 1916 prediction, on the basis of ionization potentials, that krypton and xenon fluorides should exist. However, attempts to verify that prediction — by Andreas von Antropoff and then by Otto Ruff and Walter Menzel — were unsuccessful. Ruff was one of the greatest inorganic fluorine chemists of all time, and had the required experience and experimental skills for the task at hand. Unfortunately, he pursued only argon and krypton fluorides, and not those of xenon.

In 1933, Linus Pauling published a paper predicting the existence of H_4XeO_6, and of KrF_6 and XeF_6. Using electric discharges of xenon–fluorine mixtures, his colleagues Don Yost and Albert Kaye came close to making xenon fluorides. But they did not succeed in isolating measurable amounts. That failure was taken generally as evidence that the noble gases are indeed unreactive, a principle that found its way into essentially all textbooks.

Bartlett, however, was undeterred. In 1962, at UBC, he carried out his famous experiment demonstrating that noble gases are not chemically inert. While pursuing the synthesis of platinum difluoride, PtF_2, by



UC BERKELEY

reduction of PtF_4, he purified PtF_4 by heating it in a stream of diluted fluorine in a Pyrex apparatus. He obtained a red sublimate, which he initially thought was platinum fluoride oxide, PtF_4O, but subsequently identified correctly as the ionic salt dioxygen hexafluoroplatinate, $O_2^+PtF_6^-$ — an oxidation reaction had occurred. Although the discovery of this compound was accidental, as Bartlett himself stated in Volume 9 of the *World Scientific Series in 20th Century Chemistry*, his subsequent reasoning and experiments were brilliant. He recognized that if PtF_6 can oxidize dioxygen, it should also be capable of oxidizing xenon.

His classic experiment in preparing xenon hexafluoroplatinate, $Xe^+PtF_6^-$, confirmed his reasoning, and gave rise to a worldwide interest in noble-gas chemistry. Since then, thousands of papers have been published on this subject, showing that xenon can form bonds not only to fluorine but also to many other elements of the periodic table. Moreover, noble-gas chemistry is not limited to xenon — even argon can form compounds, such as HArF. Although some of them are unstable, they nevertheless exist.

Bartlett's discovery of the first noble-gas compound was hailed by *Chemical & Engineering News* as "one of the 10 most beautiful experiments in the history of chemistry", and "one of the most important developments in inorganic chemistry in modern times". It was not, however, a one-off stroke of luck. Throughout his career, Bartlett continually demonstrated this same keen sense of reasoning and mastery as an experimentalist. He made many subsequent contributions to the field that he created, particularly in the area of xenon fluoride cations and molecular adducts of xenon fluorides with other molecules. In all of his publications, the emphasis was on quality and not quantity.

Later on, he worked on problems such as the creation of synthetic metals from graphite and graphite-like boron nitride and salts of perfluoro-aromatic cations. Another major success was his synthesis and characterization of thermodynamically unstable compounds at the limits of oxidation, such as NiF_4 and AgF_3. These compounds are powerful oxidizers and are ideal sources for the generation of high concentrations of fluorine radicals under very mild conditions.

Bartlett's achievements were recognized with 25 international and national awards, fellowships in 12 different academies and societies, and honorary degrees from 9 universities. But perhaps because of his modesty and lack of interest in lobbying for honours, he did not receive the Nobel Prize in Chemistry — which, in my opinion and those of many of his peers, was clearly deserved. This sentiment is reflected by István Hargittai in the chapter "Who did not win" of his book *The Road to Stockholm*:

> Significantly, many chemists today assume that Bartlett has won a Nobel Prize, and in this connection the most spectacular misconception can be found in Primo Levi's book, 'The Periodic Table'. On the first page of the first chapter, he mentions Bartlett's discovery: 'As late as 1962 a diligent chemist after long and ingenious efforts succeeded in forcing the Alien (xenon) to combine fleetingly with extremely avid and lively fluorine, and the feat seemed so extraordinary that he was given a Nobel prize.'

The lack of the ultimate scientific recognition, the Nobel prize, in no way diminishes the impact Bartlett has had on chemistry, both directly and through his influence on students and at scientific meetings. He was an outstanding lecturer, and at meetings invariably impressed participants with his ingenious ability to analyse problems and come up with elegant solutions. But perhaps his most memorable traits were his humbleness, friendliness, loyalty and concern for others: Neil Bartlett was not only a brilliant scholar but also a true gentleman.

Karl O. Christe
Karl O. Christe is at the Loker Research Institute, University of Southern California, Los Angeles, California 90089-1661, USA.
e-mail: kchriste@usc.edu

Arctic tropospheric warming amplification?

Arising from: R. G. Graversen, T. Mauritsen, M. Tjernström, E. Källén & G. Svensson *Nature* **451**, 53–56 (2008)

Relative rates of temperature change between the troposphere and surface, and the mechanisms that produce these changes, have long been a contentious issue. Graversen *et al.*[1], predicated upon the ERA-40 reanalysis[2], report polar tropospheric amplification of surface warming and attempt to explain this finding dynamically. Here we show (1) that data from satellites[3,4] and weather balloons[5] indicate that the ERA-40 trends are increasingly unrealistic polewards of 62° N; (2) that the two other reanalyses considered[1] exhibit very different polar trends; and (3) that the vertical profile of polar trends in ERA-40 is unrealistic, particularly above the troposphere. These quasi-independent strands of evidence imply that the pattern of warming in the Arctic troposphere is highly unlikely to be as given in ERA-40 and as reported by Graversen *et al.*[1].

Reanalyses are numerical weather-prediction systems run in hindcast mode considering all globally available observations[2]. Strenuous efforts are made to take account of both time-varying biases in the data and the impacts of the very substantially changing mix and coverage of observations. However, many aspects of the long-term behaviour of reanalyses remain unreliable[6,7] and their suitability for use in monitoring atmospheric temperature trends has been questioned by a recent expert panel[8].

Comparing ERA-40 with several observational[3–5] 'lower tropospheric' retrievals (corresponding most closely with the original analysis, peaking at about 725 hPa) over the 62.5° N to 82.5° N latitude range (Fig. 1, left-hand panels) yields good month-to-month

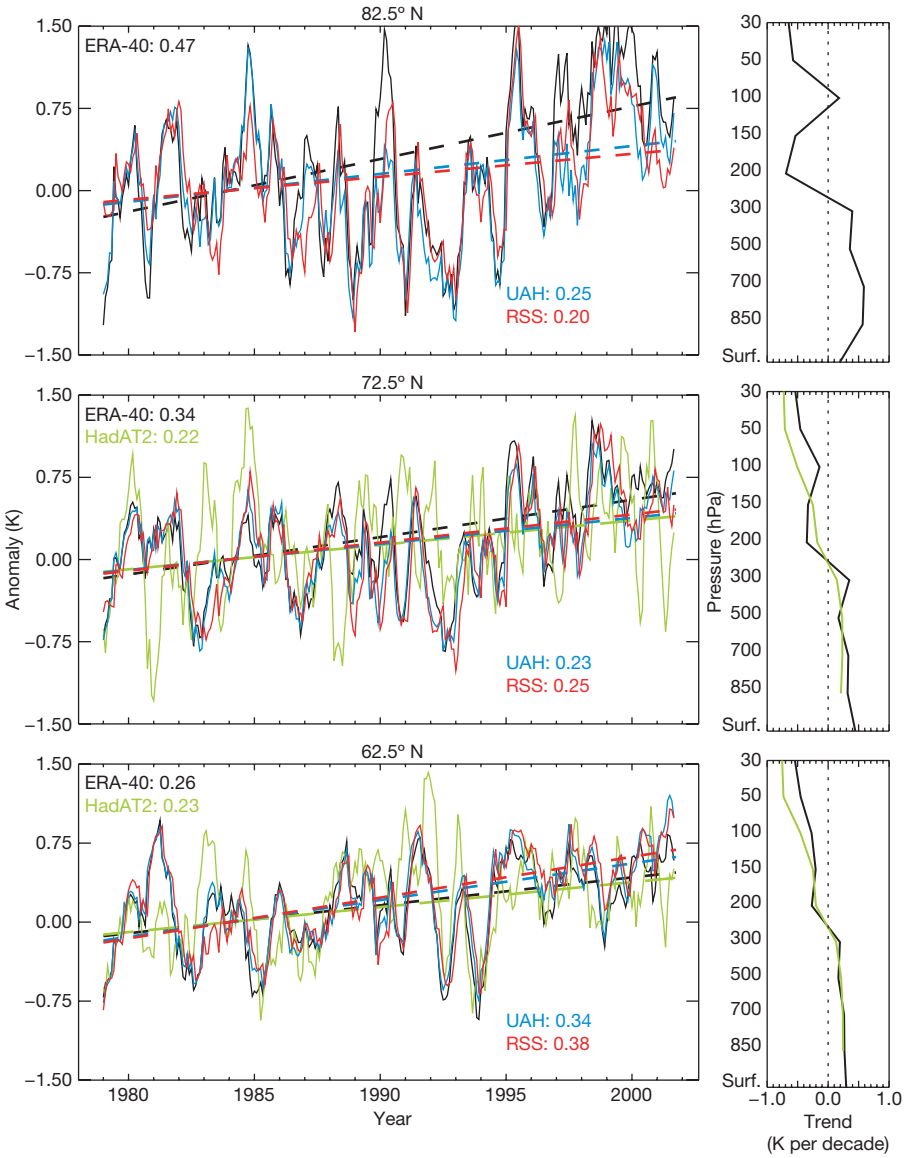


Figure 1 | Lower tropospheric retrieval data. Left-hand panels show temperature anomaly (relative to 1979–1988) monthly time series (smoothed with a simple seven-point moving filter) and trends (given as values in-line, for example ERA-40: 0.47) for three zonal bands for the broad T2LT lower tropospheric retrieval of the MSU record from UAH (ref. 3), RSS (ref. 4) and weighted equivalents from ERA-40 (ref. 2) and HadAT2 (ref. 5). Trends, calculated using a median-of-pairwise-slopes method[14], are quoted in kelvin per decade within each panel for the common period of record. Right-hand panels show vertically resolved trends on the nine HadAT2 levels for ERA-40 and HadAT2 (ref. 5). There are insufficient long-term radiosonde records at 82.5° N to assess climate trends, so there are no data here in HadAT2.

agreement, particularly with the globally complete satellite records, in accord with Graversen *et al.*[1]. Crucially, however, trends increasingly diverge as the pole is approached. High-frequency agreement is insufficient to ensure that the trend will be well characterized[9]. At 82.5° N, ERA-40 is overestimating the warming vis-à-vis available direct observational estimates by around 100%. It is north of about 80° N that ERA-40 shows the substantial warming reported by Graversen *et al.*[1]. At these latitudes, however, there are very few either conventional or space-based observations available to constrain the reanalyses. Therefore, the reality of these trends, given the lack of support from the available observational estimates[3,4] at 82.5° N, must be questioned.

Indeed, a comparison of Fig. 1 of Graversen *et al.*[1] with their Supplementary Figs 2 and 3 shows that the trend is not robust across different reanalyses systems. Although NCEP (ref. 10) can be considered a first-generation reanalysis, both ERA-40 (ref. 2) and the even newer JRA-25 (ref. 11) are second-generation reanalyses. The degree of pattern correspondence between these is visually poor, and the trend magnitudes differ substantially. This lack of robustness of the reported Arctic amplification signal implies that it is not necessarily a real-world feature.

Finally, a consideration of the full atmospheric profile rather than just that below 250 hPa shows that the ERA-40 trends become increasingly unrealistic with latitude (Fig. 1, right-hand panels). At 62.5° N, where radiosondes reporting temperatures, humidity and winds on distinct levels are plentiful, the ERA-40 trend looks realistic. Farther north, however, the availability of *in situ* radiosondes declines and the reanalysis is effectively unconstrained by *in situ* observations. Beyond 82.5° N, the reanalysis is constrained only by off-nadir views from infrared satellite observations. These are unlikely to be homogeneous. Furthermore, because they represent deep layers they cannot necessarily fully anchor the reanalysis temperatures, which may therefore have been affected by vertically differentiated model biases.

Taken together, the evidence implies that the reported Arctic tropospheric amplification is a non-climatic artefact in ERA-40. This reinforces the importance of treating any single data set, be it observational or derived, with extreme caution[12]. It does not imply that current reanalyses are unfit for the majority of purposes to which they are put. It does, however, reaffirm the importance of a properly resourced and scientifically robust attempt to create a truly climate-quality reanalysis product: a product that adequately retains long-term trend fidelity in all meteorological parameters[13].

Peter W. Thorne[1]
[1]Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK.
e-mail: peter.thorne@metoffice.gov.uk

Received 16 January; accepted 9 May 2008.

1. Graversen, R. G., Mauritsen, T., Tjernström, M., Källén, E. & Svensson, G. Vertical structure of recent Arctic warming. *Nature* **451,** 53–56 (2008).
2. Uppala, S. M. *et al.* The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.* **131,** 2961–3012 (2005).
3. Christy, J. R., Spencer, R. W., Norris, W. B., Braswell, W. D. & Parker, D. E. Error estimates of version 5.0 of MSU-AMSU bulk atmospheric temperatures. *J. Atmos. Ocean. Technol.* **20,** 613–629 (2003).
4. Mears, C. A. & Wentz, F. J. The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science* **309,** 1548–1551 (2005).
5. Thorne, P. W. *et al.* Revisiting radiosonde upper air temperatures from 1958 to 2002. *J. Geophys. Res.* **110,** D18105 (2005).
6. Mears, C. A., Santer, B. D., Wentz, F. J., Taylor, K. E. & Wehner, M. F. Relationship between temperature and precipitable water changes over tropical oceans. *Geophys. Res. Lett.* **34,** doi:10.1029/2007GL031936 (2007).
7. Bengtsson, L., Hodges, K. I. & Hagemann, S. Sensitivity of the ERA40 reanalysis to the observing system: determination of the global atmospheric circulation from reduced observations. *Tellus A* **56,** 456–471 (2004).
8. Karl, T. R., Hassol, S. J., Miller, C. D. & Murray, W. L. (eds) *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences* (Synthesis and Assessment Product 1.1, US Climate Change Science Program, 2006).
9. Sherwood, S. C., Titchner, H. A., Thorne, P. W. & McCarthy, M. C. How do we tell which estimates of past climate change are correct? *Int. J. Climatol.* (submitted).
10. Kalnay, E. *et al.* The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77,** 437–470 (1996).
11. Onogi, K. *et al.* The JRA-25 reanalysis. *J. Meteorol. Soc. Jpn* **85,** 369–432 (2007).
12. Thorne, P. W., Parker, D. E., Christy, J. R. & Mears, C. A. Uncertainties in climate trends - Lessons from upper-air temperature records. *Bull. Am. Meteorol. Soc.* **86,** 1437–1442 (2005).
13. Bengtsson, L. *et al.* The need for a dynamical climate reanalysis. *Bull. Am. Meteorol. Soc.* **88,** 495–501 (2007).
14. Lanzante, J. R. Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data. *Int. J. Climatol.* **16,** 1197–1226 (1996).

doi:10.1038/nature07256

Recent Arctic warming vertical structure contested

Arising from: R. G. Graversen, T. Mauritsen, M. Tjernström, E. Källén & G. Svensson *Nature* 451, 53–56 (2008)

The vertical structure of the recent Arctic warming contains information about the processes governing Arctic climate trends. Graversen *et al.* argue[1], on the basis of ERA-40 reanalysis[2] data, that a distinct maximum in 1979–2001 warm-season (April–October) Arctic temperature trends appears around 3 km above ground. Here we show that this is due to the heterogeneous nature of the data source, which incorporates information from satellites and radiosondes. Radiosonde data alone suggest the warming was strongest near ground.

Graversen *et al.*[1] claim that the warm-season temperature trend has a maximum at around 700 hPa, polewards of 75° N, and argue that anomalous heat advection from more southerly latitudes is important. However, the ERA-40 reanalysis may not be suitable for trend analysis as it incorporates information from different observing systems such as satellite and radiosonde, which might be inconsistent, in particular with respect to trends[3,4]. Radiosonde measurements provide vertically resolved temperature profiles in the troposphere, whereas satellites provide information on a weighted average over a thick layer. Furthermore, the ERA-40 assimilation system extrapolates information from data-rich to data-sparse areas, which is less reliable than observations. The ERA-40 reanalysis in the polar region has not been sufficiently validated by *in situ* observations and documented[2,5] problems with satellite radiance assimilations over the Arctic Ocean could lead to spurious trends.

A map of warm-season trends at 700 hPa (the peak level of the polar warming trend in ref. 1) from ERA-40 and radiosonde observations[6,7] confirms that the enhanced warming signal lies mostly in areas with no radiosonde data coverage (Fig. 1a). This is particularly so polewards of 75° N, where the trend appears strongest in ref. 1. Moreover, the few radiosonde data available near or polewards of 75° N show modest trends. To illustrate the effects on the vertical structure of the trend, we calculated zonally averaged vertical temperature trends from (1) ERA-40 reanalysis data, (2) such data subsampled to locations where radiosonde information is available (that is, where ERA-40 is best constrained) and (3) from only radiosonde data. The trend in the reanalysis (Fig. 1b) is a reproduction of Fig. 4a in ref. 1 and exhibits a maximum at 700 hPa, polewards of 75° N. Subsampling the ERA-40 reanalysis (Fig. 1c) reveals clearly different trends, and calculating trends directly from radiosondes alters matters even further (Fig. 1d). The result is independent of the methods used to homogenize the radiosonde data (unadjusted, RAOBCORE v1.4 (ref. 7) and RICH (ref. 8); not shown). The radiosonde data (note that some regions are not well covered and some levels are

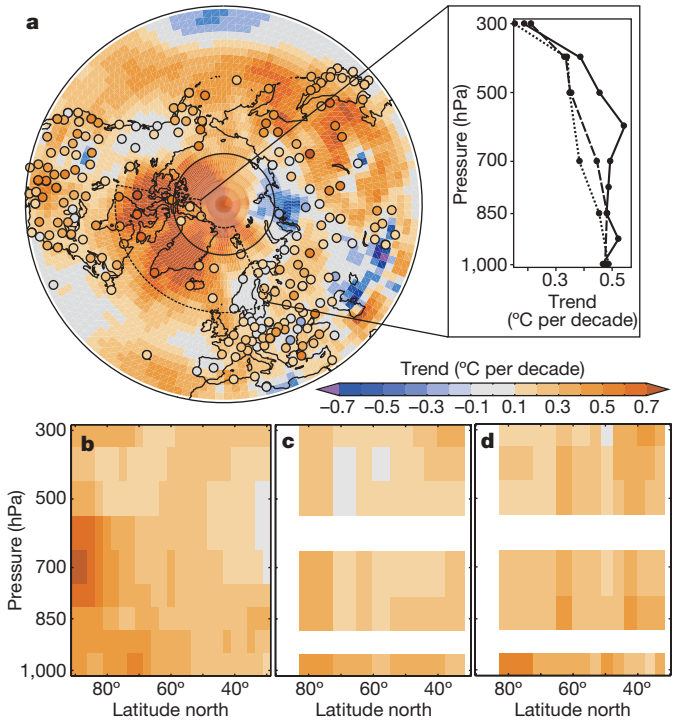


Figure 1 | Vertical structure of Arctic temperature trends for April to October, 1979–2001. Trends were calculated from seasonally averaged monthly anomalies using least-squares regression (not more than one missing month per season allowed, not more than five missing seasons in 1979–2001, neither first nor last two years can be missing). **a**, Trend field at 700 hPa from ERA-40 (ref. 2) and from radiosonde data[6,7] (circles) with 75° N latitude circle indicated by the thin solid line. **b**, Trends of zonal-mean temperature as a function of latitude and altitude from ERA-40. **c**, Same as **b**, but from ERA-40 subsampled to the locations and times where radiosonde data are available (anomalies zonally averaged in equal-area latitude bands). **d**, Same as **c**, but for radiosonde data. Inset in **a**, average trend profiles of the region 58° N–82° N, 100° W–25° E for full ERA-40 (solid line), subsampled ERA-40 (dashed) and radiosonde data (dotted).

missing because of inconsistent reporting) have their strongest trend near the ground, not above the boundary layer as in the full reanalysis. This is important because boundary layer processes are much more locally driven and simultaneously not well represented in a reanalysis. The same result is found when analysing a subregion with relatively even radiosonde coverage (inset, Fig. 1a), and during the remainder of the year (not shown).

Arctic climate is controlled by processes operating on scales from local to global, including transport effects; forcings such as greenhouse gases, aerosols and clouds; and feedbacks such as the well-known sea-ice–albedo feedback. The temperature profile can be a clue to the underlying processes, but to disentangle the contributions to Arctic temperature trends fully, vertical temperature structures should be addressed in a regionally and seasonally resolved manner. Furthermore, the large interannual variability in the Arctic, coupled with the sensitivity of trends to both end points and season definitions, suggests care should be taken in interpreting trends over short periods.

In conclusion, some features of the temperature trends calculated in ref. 1 reflect possible inhomogeneities or artefacts in the ERA-40 reanalysis rather than true climate signals, as they appear not to be supported by observations. ERA-40 reanalysis is a valuable tool in calculating circulation effects, especially on a subdecadal basis, but inhomogeneities and gaps in the global observing system tend to make trends from reanalyses unreliable, particularly in data-sparse regions.

A. N. Grant[1], **S. Brönnimann**[1] **& L. Haimberger**[2]
[1]Institute for Atmospheric and Climate Science, ETH Zurich, Universitätstrasse 16, CH-8092 Zurich, Switzerland.
e-mail: andrea.grant@env.ethz.ch
[2]Department of Meteorology and Geophysics, University of Vienna, Althanstrasse 14, A-1090 Vienna, Austria.

Received 23 January; accepted 12 May 2008.

1. Graversen, R. G., Mauritsen, T., Tjernström, M., Källén, E. & Svensson, G. Vertical structure of recent Arctic warming. *Nature* **451**, 53–56 (2008).
2. Uppala, S. M. *et al.* The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.* **131**, 2961–3012 (2005).
3. Fu, Q., Johanson, C. M., Warren, S. G. & Seidel, D. J. Contribution of stratospheric cooling to satellite-inferred tropospheric temperature trends. *Nature* **429**, 55–58 (2004).
4. Mears, C. A. & Wentz, F. J. The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science* **309**, 1548–1551 (2005).
5. Bromwich, D. & Wang, S.-H. Evaluation of the NCEP-NCAR and ECMWF 15- and 40-yr reanalyses using rawindsonde data from two independent Arctic field experiments. *Mon. Weath. Rev.* **133**, 3562–3578 (2005).
6. Durre, I., Vose, R. & Wuertz, D. Overview of the integrated global radiosonde archive. *J. Clim.* **19**, 53–68 (2006).
7. Haimberger, L. Homogenization of radiosonde temperature time series using innovation statistics. *J. Clim.* **20**, 1377–1403 (2007).
8. Haimberger, L. *et al.* Towards elimination of the warm bias in historic radiosonde temperature records — some new results from a comprehensive intercomparison of upper air data. *J. Clim.* (in the press).

doi:10.1038/nature07257

Arctic warming aloft is data set dependent

Arising from: R. G. Graversen, T. Mauritsen, M. Tjernström, E. Källén & G. Svensson *Nature* 451, 53–56 (2008)

Arctic sea ice and snow on land have retreated polewards at an alarming pace in the past few decades[1]. Such retreat locally amplifies surface warming through a positive feedback, which causes the Arctic surface to warm faster than the rest of the globe. In contrast, ice and snow retreat causes little warming in the atmosphere above when the stable winter atmosphere inhibits vertical heat exchange. We therefore find surprising the recent report by Graversen *et al.*[2] in which they claim that recent Arctic atmospheric warming extends far deeper into the atmosphere than expected, and can even exceed the surface warming during the polar night. Using a different data set, we show that there is much less warming aloft in winter, consistent with the recent retreat of ice and snow, as well as recent changes in atmospheric heat transport.

Graversen *et al.*[2] compute trends for 1979–2001 from ERA-40 reanalysis, which is a hybrid product using many types of raw observational data assimilated with a consistent global analysis system. The assimilation compensates for some but not all of the variations in the observing system over time that may compromise the veracity of the temperature trend analyses[3]. Figure 1 compares temperature trends in winter from the ERA-40 reanalysis with climate-quality records from satellite observations[4]. Trends in the Arctic winter aloft are strongly data set dependent: the observed trend is 75% less than reanalysis in the middle troposphere and 40% less than in the lower-middle troposphere. In comparison with the observations, the reanalysis exaggerates polar amplification aloft by overestimating the Arctic atmospheric warming and underestimating the Northern

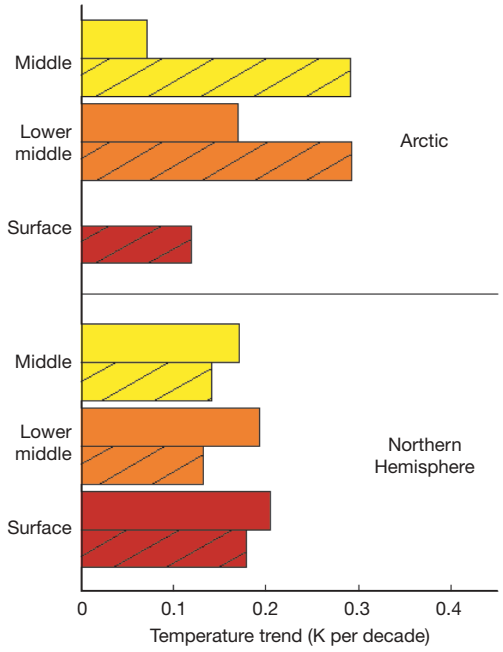


Figure 1 | Temperature trends over the Northern Hemisphere (0°–82.5° N) and the Arctic (65° N–82.5° N) for 1979–2001. Trends are for temperatures at the surface and in the lower-middle and middle troposphere from the ERA-40 reanalysis (hatched) and observations (solid) in the winter (December–February) season. The observed trends are derived from the HadCrut3v (ref. 8) data set for the surface temperature and from a satellite microwave sounding unit[9] (MSU; RSS version 3) for the temperatures in the lower-middle[10] and middle[11,12] troposphere. Observed surface temperatures in the Arctic are not shown, because they are spatially incomplete. For a direct comparison with the MSU observations, synthetic temperatures in the lower-middle and middle troposphere are computed from the ERA-40 reanalysis by applying the MSU weighting functions[3].

Hemisphere atmospheric warming in every season. Specifically, for trends in annual means in the reanalysis for 1979–2001, the Arctic warms 2.7 times more than the Northern Hemisphere in the lower-middle troposphere, in comparison with just 1.5 times more in the observations.

During the polar night, solar absorption at the surface is absent or weak. At the same time, the atmosphere transports a substantial amount of heat northwards from lower latitudes, with heating rates in the Arctic that maximize at about 1,500 m in winter[5]. For these reasons and others, strong radiative cooling at the surface causes frequent lower-tropospheric temperature inversions, which are very stable and damp vertical heat transfer during the polar night. When ice and snow retreat, some of the heat from increased solar absorption is stored at the ocean surface and is released during the cold seasons without warming the atmosphere aloft very much.

It has been concluded that northwards atmospheric heat transport into the Arctic should increase in a warming world[6,7] owing to increased evaporation in the tropics and subsequent condensation in the high latitudes. This increase in latent heat transport is somewhat counterbalanced by a decrease in sensible heat transport, as Arctic amplification decreases the pole-to-equator temperature gradient. Models indicate that warming aloft would not outpace the surface warming after considering increased northwards atmospheric heat transport along with the retreat of ice and snow[7]. Graversen *et al.*[2] find that the change in northwards atmospheric heat transport is not a substantial source of heating aloft in midwinter (January–February) in the Arctic.

The smaller warming trends aloft in the observations in winter are more consistent with the amplification of surface warming from ice and snow retreat and the lack of change in the northwards atmospheric heat transport for 1979–2001. This consistent set of observations calls into question the results of Graversen *et al.*[2] obtained for the polar night.

Cecilia M. Bitz[1] & Qiang Fu[1]

[1]Atmospheric Science Department, 408 Atmospheric Sciences and Geophysics Hall, University of Washington, Seattle, Washington 98195, USA.
e-mail: bitz@atmos.washington.edu

Received 5 February; accepted 9 May 2008.

1. Lemke, P. *et al.* in *Climate Change 2007: The Physical Science Basis* (eds Solomon, S. *et al.*) 337–383 (Contribution of Working Group I to the Fourth Assessment Report of the IPCC, Cambridge Univ. Press, 2007).
2. Graversen, R. G., Mauritsen, T., Tjernström, M., Källén, E. & Svensson, G. Vertical structure of recent Arctic warming. *Nature* **451,** 53–56 (2008).
3. Johanson, C. M. & Fu, Q. Antarctic atmospheric temperature trend patterns from satellite observations. *Geophys. Res. Lett.* **34,** doi:10.1029/2006GL029108 (2007).
4. Karl, T. R., Hassol, S. J., Miller, C. D. & Murray, W. L. (eds) *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences* (Synthesis and Assessment Product 1.1, US Climate Change Science Program, 2006).
5. Overland, J. E. & Turret, P. in *The Polar Oceans and Their Role in Shaping the Global Environment* (eds Johannessen, O. M., Muench, R. & Overland, J. E.) 313–325 (American Geophysical Union, 1994).
6. Alexseev, V. A., Langen, P. L. & Bates, J. R. Polar amplification of surface warming on an aquaplanet in "ghost forcing" experiments without sea ice feedbacks. *Clim. Dyn.* **24,** 655–666 (2005).
7. Cai, M. & Lu, J. Dynamical greenhouse-plus feedback and polar warming amplification. Part II: meridional and vertical asymmetries of the global warming. *Clim. Dyn.* **29,** 375–391 (2007).
8. Brohan, P., Kennedy, J. J., Haris, I., Tett, S. F. B. & Jones, P. D. Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *J. Geophys. Res.* **111,** doi:10.1029/2005JD006548 (2006).
9. Mears, C. A., Schabel, M. C. & Wentz, F. J. A reanalysis of the MSU channel 2 tropospheric temperature record. *J. Clim.* **16,** 3650–3664 (2003).
10. Mears, C. A. & Wentz, F. J. The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science* **309,** 1548–1551 (2005).
11. Fu, Q., Johanson, C. M., Warren, S. G. & Seidel, D. J. Contribution of stratospheric cooling to satellite-inferred tropospheric temperature trends. *Nature* **429,** 55–58 (2004).
12. Johanson, C. M. & Fu, Q. Robustness of tropospheric temperature trends from MSU channels 2 and 4. *J. Clim.* **19,** 4234–4242 (2006).

doi:10.1038/nature07258

Graversen et al. *reply*

Replying to: P. W. Thorne *Nature* 455, doi:10.1038/nature07256; A. N. Grant, S. Brönnimann & L. Haimberger *Nature* 455, doi:10.1038/nature07257; C. M. Bitz & Q. Fu *Nature* 455, doi:10.1038/nature07258 (2008)

These three communications[1–3] question the validity of some of our conclusions[4]. We found Arctic temperature trend amplification well above the boundary layer. In summer, the maximum amplification is found at a height of around 2 km, and no amplification is encountered near the surface. These findings appear in two state-of-the-art reanalyses, ERA-40 (ref. 5) and JRA-25 (ref. 6). Both these data sets show roughly the same overall vertical structure, and we believe our conclusions can be based on either of them. However, they show

considerable differences regarding the magnitudes of the Arctic trends (see our Supplementary Information[4]), but our conclusions are not based on the absolute magnitudes.

A reanalysis synthesizes all available observations and uses a physically based model of the atmosphere to weigh the observations against each other and to extrapolate the observed information in space and time to unobserved parts of the atmosphere. The assimilation procedure takes observational as well as model uncertainties into account. In ERA-40 and JRA-25, the strongest observational constraint on the Arctic temperatures aloft is provided by assimilation of satellite observations, such as microwave sounding unit (MSU) radiances, as *in situ* observations are few in this region. In the assimilation process, careful bias adjustment has been applied to the satellite observations[5].

We examined the agreement of the MSU satellite observations (RSS analysis[7] TLT v3.1 and TMT v3.2) with the vertical structure of ERA-40 and JRA-25. Arctic amplification is encountered in the channel representing the lower troposphere. In summer, both the lower-troposphere and the middle-troposphere channels indicate considerable warming over the Arctic. Because the Arctic surface temperature is constrained to be close to the melting point during this season, this warming must occur aloft, in accordance with the two reanalyses.

The annual lower-troposphere MSU trend reaches 0.46 K per decade at 81.25° N, calculated on the basis of a least-squares fit. We therefore find it surprising that Thorne[1] estimates a high-latitude trend of only 0.2 K per decade. Bitz and Fu[3] report winter trends in the lower troposphere of around 0.2 K per decade both for the Arctic and the Northern Hemisphere, which we also find. However, they report middle-troposphere trends of around 0.09 and 0.18 K per decade for the Arctic and the Northern Hemisphere, respectively. We find, on the other hand, 0.14 and 0.09 K per decade for the Arctic and the Northern Hemisphere, respectively. Hence, the MSU data show winter Arctic amplification in agreement with ERA-40 and JRA-25.

In their last paragraph, Bitz and Fu[3] indicate that ERA-40 exaggerates winter trends aloft. This might be the case; JRA-25 shows considerably smaller trends. However, our point is that, even in JRA-25, winter trends above the boundary layer are comparable to those near the surface and can hardly be linked to surface processes alone. Grant *et al.*[2] compare ERA-40 data with radiosonde observations, which are few in the Arctic. Although these observations cannot confirm the April–October warming aloft found in ERA-40, in general they show good agreement with the ERA-40 data at the points where radiosondes are available.

There is no doubt that more *in situ* observations in the Arctic are needed to enhance the quality of future reanalyses. Given the absence of such observations in historical archives, we feel that a reanalysis is likely to provide a better representation of the true state of the Arctic atmosphere than any single inhomogeneous set of a specific observation type. Satellite observations must be bias corrected and radio soundings exist almost only in the southern part of the Arctic. In a reanalysis, both of these shortcomings are consistently handled in the framework of a dynamical, global model of the atmosphere. We have given an estimate of the uncertainty associated with reanalysis data by displaying results from two different, second-generation reanalyses. Within the limits of this uncertainty we believe that our conclusions remain valid.

R. G. Graversen[1], T. Mauritsen[1], M. Tjernström[1], E. Källén[1] & G. Svensson[1]

[1]Department of Meteorology, Stockholm University, S-106 91 Stockholm, Sweden.

e-mail: rune@misu.su.se

1. Thorne, P. W. Arctic tropospheric warming amplification? *Nature* **455,** doi:10.1038/nature07256 (2008).
2. Grant, A. N., Brönnimann, S. & Haimberger, L. Recent Arctic warming vertical structure contested. *Nature* **455,** doi:10.1038/nature07257 (2008).
3. Bitz, C. M. & Fu, Q. Arctic warming aloft is data set dependent. *Nature* **455,** doi:10.1038/nature07258 (2008).
4. Graversen, R. G., Mauritsen, T., Tjernström, M., Källén, E. & Svensson, G. Vertical structure of recent Arctic warming. *Nature* **451,** 53–56 (2008).
5. Uppala, S. M. *et al.* The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.* **131,** 2961–3012 (2005).
6. Onogi, K. *et al.* The JRA-25 reanalysis. *J. Meteorol. Soc. Jpn* **85,** 369–432 (2007).
7. Mears, C. A. & Wentz, F. J. The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science* **309,** 1548–1551 (2005).

doi:10.1038/nature07259

Arctic tropospheric warming amplification?

Arising from: R. G. Graversen, T. Mauritsen, M. Tjernström, E. Källén & G. Svensson *Nature* **451**, 53–56 (2008)

Relative rates of temperature change between the troposphere and surface, and the mechanisms that produce these changes, have long been a contentious issue. Graversen *et al.*[1], predicated upon the ERA-40 reanalysis[2], report polar tropospheric amplification of surface warming and attempt to explain this finding dynamically. Here we show (1) that data from satellites[3,4] and weather balloons[5] indicate that the ERA-40 trends are increasingly unrealistic polewards of 62° N; (2) that the two other reanalyses considered[1] exhibit very different polar trends; and (3) that the vertical profile of polar trends in ERA-40 is unrealistic, particularly above the troposphere. These quasi-independent strands of evidence imply that the pattern of warming in the Arctic troposphere is highly unlikely to be as given in ERA-40 and as reported by Graversen *et al.*[1].

Reanalyses are numerical weather-prediction systems run in hindcast mode considering all globally available observations[2]. Strenuous efforts are made to take account of both time-varying biases in the data and the impacts of the very substantially changing mix and coverage of observations. However, many aspects of the long-term behaviour of reanalyses remain unreliable[6,7] and their suitability for use in monitoring atmospheric temperature trends has been questioned by a recent expert panel[8].

Comparing ERA-40 with several observational[3–5] 'lower tropospheric' retrievals (corresponding most closely with the original analysis, peaking at about 725 hPa) over the 62.5° N to 82.5° N latitude range (Fig. 1, left-hand panels) yields good month-to-month

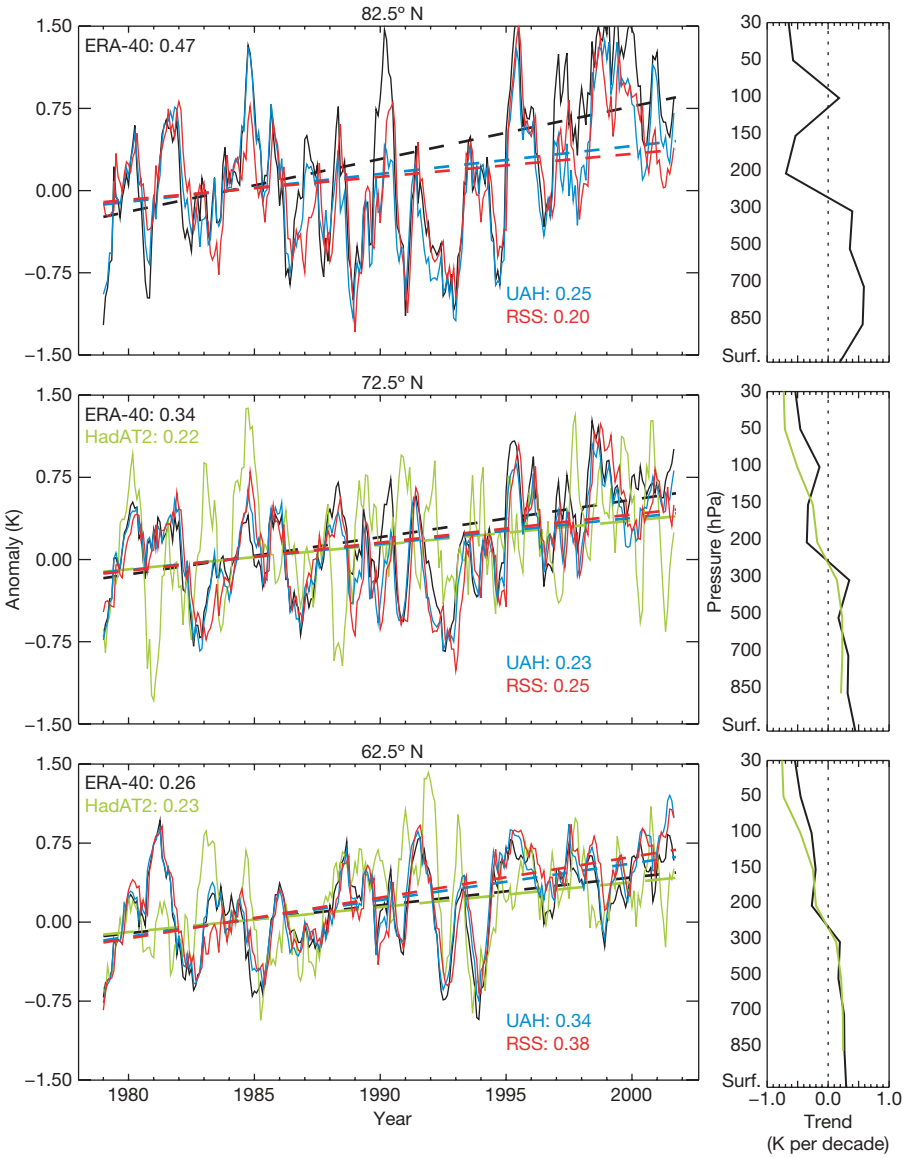


Figure 1 | Lower tropospheric retrieval data. Left-hand panels show temperature anomaly (relative to 1979–1988) monthly time series (smoothed with a simple seven-point moving filter) and trends (given as values in-line, for example ERA-40: 0.47) for three zonal bands for the broad T2LT lower tropospheric retrieval of the MSU record from UAH (ref. 3), RSS (ref. 4) and weighted equivalents from ERA-40 (ref. 2) and HadAT2 (ref. 5). Trends, calculated using a median-of-pairwise-slopes method[14], are quoted in kelvin per decade within each panel for the common period of record. Right-hand panels show vertically resolved trends on the nine HadAT2 levels for ERA-40 and HadAT2 (ref. 5). There are insufficient long-term radiosonde records at 82.5° N to assess climate trends, so there are no data here in HadAT2.

agreement, particularly with the globally complete satellite records, in accord with Graversen *et al.*[1]. Crucially, however, trends increasingly diverge as the pole is approached. High-frequency agreement is insufficient to ensure that the trend will be well characterized[9]. At 82.5° N, ERA-40 is overestimating the warming vis-à-vis available direct observational estimates by around 100%. It is north of about 80° N that ERA-40 shows the substantial warming reported by Graversen *et al.*[1]. At these latitudes, however, there are very few either conventional or space-based observations available to constrain the reanalyses. Therefore, the reality of these trends, given the lack of support from the available observational estimates[3,4] at 82.5° N, must be questioned.

Indeed, a comparison of Fig. 1 of Graversen *et al.*[1] with their Supplementary Figs 2 and 3 shows that the trend is not robust across different reanalyses systems. Although NCEP (ref. 10) can be considered a first-generation reanalysis, both ERA-40 (ref. 2) and the even newer JRA-25 (ref. 11) are second-generation reanalyses. The degree of pattern correspondence between these is visually poor, and the trend magnitudes differ substantially. This lack of robustness of the reported Arctic amplification signal implies that it is not necessarily a real-world feature.

Finally, a consideration of the full atmospheric profile rather than just that below 250 hPa shows that the ERA-40 trends become increasingly unrealistic with latitude (Fig. 1, right-hand panels). At 62.5° N, where radiosondes reporting temperatures, humidity and winds on distinct levels are plentiful, the ERA-40 trend looks realistic. Farther north, however, the availability of *in situ* radiosondes declines and the reanalysis is effectively unconstrained by *in situ* observations. Beyond 82.5° N, the reanalysis is constrained only by off-nadir views from infrared satellite observations. These are unlikely to be homogeneous. Furthermore, because they represent deep layers they cannot necessarily fully anchor the reanalysis temperatures, which may therefore have been affected by vertically differentiated model biases.

Taken together, the evidence implies that the reported Arctic tropospheric amplification is a non-climatic artefact in ERA-40. This reinforces the importance of treating any single data set, be it observational or derived, with extreme caution[12]. It does not imply that current reanalyses are unfit for the majority of purposes to which they are put. It does, however, reaffirm the importance of a properly resourced and scientifically robust attempt to create a truly climate-quality reanalysis product: a product that adequately retains long-term trend fidelity in all meteorological parameters[13].

Peter W. Thorne[1]
[1]Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK.
e-mail: peter.thorne@metoffice.gov.uk

Received 16 January; accepted 9 May 2008.

1. Graversen, R. G., Mauritsen, T., Tjernström, M., Källén, E. & Svensson, G. Vertical structure of recent Arctic warming. *Nature* **451,** 53–56 (2008).
2. Uppala, S. M. *et al.* The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.* **131,** 2961–3012 (2005).
3. Christy, J. R., Spencer, R. W., Norris, W. B., Braswell, W. D. & Parker, D. E. Error estimates of version 5.0 of MSU-AMSU bulk atmospheric temperatures. *J. Atmos. Ocean. Technol.* **20,** 613–629 (2003).
4. Mears, C. A. & Wentz, F. J. The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science* **309,** 1548–1551 (2005).
5. Thorne, P. W. *et al.* Revisiting radiosonde upper air temperatures from 1958 to 2002. *J. Geophys. Res.* **110,** D18105 (2005).
6. Mears, C. A., Santer, B. D., Wentz, F. J., Taylor, K. E. & Wehner, M. F. Relationship between temperature and precipitable water changes over tropical oceans. *Geophys. Res. Lett.* **34,** doi:10.1029/2007GL031936 (2007).
7. Bengtsson, L., Hodges, K. I. & Hagemann, S. Sensitivity of the ERA40 reanalysis to the observing system: determination of the global atmospheric circulation from reduced observations. *Tellus A* **56,** 456–471 (2004).
8. Karl, T. R., Hassol, S. J., Miller, C. D. & Murray, W. L. (eds) *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences* (Synthesis and Assessment Product 1.1, US Climate Change Science Program, 2006).
9. Sherwood, S. C., Titchner, H. A., Thorne, P. W. & McCarthy, M. C. How do we tell which estimates of past climate change are correct? *Int. J. Climatol.* (submitted).
10. Kalnay, E. *et al.* The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77,** 437–470 (1996).
11. Onogi, K. *et al.* The JRA-25 reanalysis. *J. Meteorol. Soc. Jpn* **85,** 369–432 (2007).
12. Thorne, P. W., Parker, D. E., Christy, J. R. & Mears, C. A. Uncertainties in climate trends - Lessons from upper-air temperature records. *Bull. Am. Meteorol. Soc.* **86,** 1437–1442 (2005).
13. Bengtsson, L. *et al.* The need for a dynamical climate reanalysis. *Bull. Am. Meteorol. Soc.* **88,** 495–501 (2007).
14. Lanzante, J. R. Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data. *Int. J. Climatol.* **16,** 1197–1226 (1996).

doi:10.1038/nature07256

Recent Arctic warming vertical structure contested

Arising from: R. G. Graversen, T. Mauritsen, M. Tjernström, E. Källén & G. Svensson *Nature* 451, 53–56 (2008)

The vertical structure of the recent Arctic warming contains information about the processes governing Arctic climate trends. Graversen *et al.* argue[1], on the basis of ERA-40 reanalysis[2] data, that a distinct maximum in 1979–2001 warm-season (April–October) Arctic temperature trends appears around 3 km above ground. Here we show that this is due to the heterogeneous nature of the data source, which incorporates information from satellites and radiosondes. Radiosonde data alone suggest the warming was strongest near ground.

Graversen *et al.*[1] claim that the warm-season temperature trend has a maximum at around 700 hPa, polewards of 75° N, and argue that anomalous heat advection from more southerly latitudes is important. However, the ERA-40 reanalysis may not be suitable for trend analysis as it incorporates information from different observing systems such as satellite and radiosonde, which might be inconsistent, in particular with respect to trends[3,4]. Radiosonde measurements provide vertically resolved temperature profiles in the troposphere, whereas satellites provide information on a weighted average over a thick layer. Furthermore, the ERA-40 assimilation system extrapolates information from data-rich to data-sparse areas, which is less reliable than observations. The ERA-40 reanalysis in the polar region has not been sufficiently validated by *in situ* observations and documented[2,5] problems with satellite radiance assimilations over the Arctic Ocean could lead to spurious trends.

A map of warm-season trends at 700 hPa (the peak level of the polar warming trend in ref. 1) from ERA-40 and radiosonde observations[6,7] confirms that the enhanced warming signal lies mostly in areas with no radiosonde data coverage (Fig. 1a). This is particularly so polewards of 75° N, where the trend appears strongest in ref. 1. Moreover, the few radiosonde data available near or polewards of 75° N show modest trends. To illustrate the effects on the vertical structure of the trend, we calculated zonally averaged vertical temperature trends from (1) ERA-40 reanalysis data, (2) such data subsampled to locations where radiosonde information is available (that is, where ERA-40 is best constrained) and (3) from only radiosonde data. The trend in the reanalysis (Fig. 1b) is a reproduction of Fig. 4a in ref. 1 and exhibits a maximum at 700 hPa, polewards of 75° N. Subsampling the ERA-40 reanalysis (Fig. 1c) reveals clearly different trends, and calculating trends directly from radiosondes alters matters even further (Fig. 1d). The result is independent of the methods used to homogenize the radiosonde data (unadjusted, RAOBCORE v1.4 (ref. 7) and RICH (ref. 8); not shown). The radiosonde data (note that some regions are not well covered and some levels are

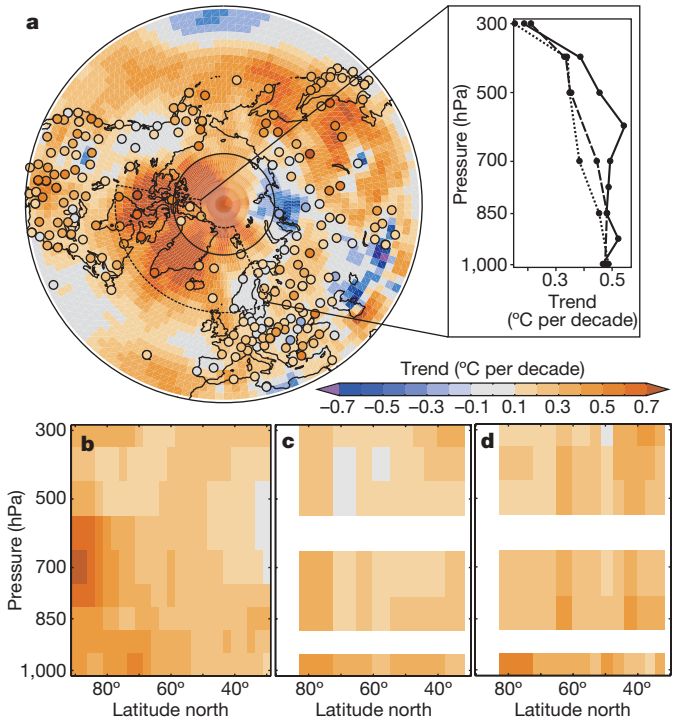


Figure 1 | Vertical structure of Arctic temperature trends for April to October, 1979–2001. Trends were calculated from seasonally averaged monthly anomalies using least-squares regression (not more than one missing month per season allowed, not more than five missing seasons in 1979–2001, neither first nor last two years can be missing). **a**, Trend field at 700 hPa from ERA-40 (ref. 2) and from radiosonde data[6,7] (circles) with 75° N latitude circle indicated by the thin solid line. **b**, Trends of zonal-mean temperature as a function of latitude and altitude from ERA-40. **c**, Same as **b**, but from ERA-40 subsampled to the locations and times where radiosonde data are available (anomalies zonally averaged in equal-area latitude bands). **d**, Same as **c**, but for radiosonde data. Inset in **a**, average trend profiles of the region 58° N–82° N, 100° W–25° E for full ERA-40 (solid line), subsampled ERA-40 (dashed) and radiosonde data (dotted).

missing because of inconsistent reporting) have their strongest trend near the ground, not above the boundary layer as in the full reanalysis. This is important because boundary layer processes are much more locally driven and simultaneously not well represented in a reanalysis. The same result is found when analysing a subregion with relatively even radiosonde coverage (inset, Fig. 1a), and during the remainder of the year (not shown).

Arctic climate is controlled by processes operating on scales from local to global, including transport effects; forcings such as greenhouse gases, aerosols and clouds; and feedbacks such as the well-known sea-ice–albedo feedback. The temperature profile can be a clue to the underlying processes, but to disentangle the contributions to Arctic temperature trends fully, vertical temperature structures should be addressed in a regionally and seasonally resolved manner. Furthermore, the large interannual variability in the Arctic, coupled with the sensitivity of trends to both end points and season definitions, suggests care should be taken in interpreting trends over short periods.

In conclusion, some features of the temperature trends calculated in ref. 1 reflect possible inhomogeneities or artefacts in the ERA-40 reanalysis rather than true climate signals, as they appear not to be supported by observations. ERA-40 reanalysis is a valuable tool in calculating circulation effects, especially on a subdecadal basis, but inhomogeneities and gaps in the global observing system tend to make trends from reanalyses unreliable, particularly in data-sparse regions.

A. N. Grant[1], **S. Brönnimann**[1] **& L. Haimberger**[2]
[1]Institute for Atmospheric and Climate Science, ETH Zurich, Universitätstrasse 16, CH-8092 Zurich, Switzerland.
e-mail: andrea.grant@env.ethz.ch
[2]Department of Meteorology and Geophysics, University of Vienna, Althanstrasse 14, A-1090 Vienna, Austria.

Received 23 January; accepted 12 May 2008.

1. Graversen, R. G., Mauritsen, T., Tjernström, M., Källén, E. & Svensson, G. Vertical structure of recent Arctic warming. *Nature* **451**, 53–56 (2008).
2. Uppala, S. M. *et al.* The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.* **131**, 2961–3012 (2005).
3. Fu, Q., Johanson, C. M., Warren, S. G. & Seidel, D. J. Contribution of stratospheric cooling to satellite-inferred tropospheric temperature trends. *Nature* **429**, 55–58 (2004).
4. Mears, C. A. & Wentz, F. J. The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science* **309**, 1548–1551 (2005).
5. Bromwich, D. & Wang, S.-H. Evaluation of the NCEP-NCAR and ECMWF 15- and 40-yr reanalyses using rawindsonde data from two independent Arctic field experiments. *Mon. Weath. Rev.* **133**, 3562–3578 (2005).
6. Durre, I., Vose, R. & Wuertz, D. Overview of the integrated global radiosonde archive. *J. Clim.* **19**, 53–68 (2006).
7. Haimberger, L. Homogenization of radiosonde temperature time series using innovation statistics. *J. Clim.* **20**, 1377–1403 (2007).
8. Haimberger, L. *et al.* Towards elimination of the warm bias in historic radiosonde temperature records — some new results from a comprehensive intercomparison of upper air data. *J. Clim.* (in the press).

doi:10.1038/nature07257

Arctic warming aloft is data set dependent

Arising from: R. G. Graversen, T. Mauritsen, M. Tjernström, E. Källén & G. Svensson *Nature* 451, 53–56 (2008)

Arctic sea ice and snow on land have retreated polewards at an alarming pace in the past few decades[1]. Such retreat locally amplifies surface warming through a positive feedback, which causes the Arctic surface to warm faster than the rest of the globe. In contrast, ice and snow retreat causes little warming in the atmosphere above when the stable winter atmosphere inhibits vertical heat exchange. We therefore find surprising the recent report by Graversen *et al.*[2] in which they claim that recent Arctic atmospheric warming extends far deeper into the atmosphere than expected, and can even exceed the surface warming during the polar night. Using a different data set, we show that there is much less warming aloft in winter, consistent with the recent retreat of ice and snow, as well as recent changes in atmospheric heat transport.

Graversen *et al.*[2] compute trends for 1979–2001 from ERA-40 reanalysis, which is a hybrid product using many types of raw observational data assimilated with a consistent global analysis system. The assimilation compensates for some but not all of the variations in the observing system over time that may compromise the veracity of the temperature trend analyses[3]. Figure 1 compares temperature trends in winter from the ERA-40 reanalysis with climate-quality records from satellite observations[4]. Trends in the Arctic winter aloft are strongly data set dependent: the observed trend is 75% less than reanalysis in the middle troposphere and 40% less than in the lower-middle troposphere. In comparison with the observations, the reanalysis exaggerates polar amplification aloft by overestimating the Arctic atmospheric warming and underestimating the Northern

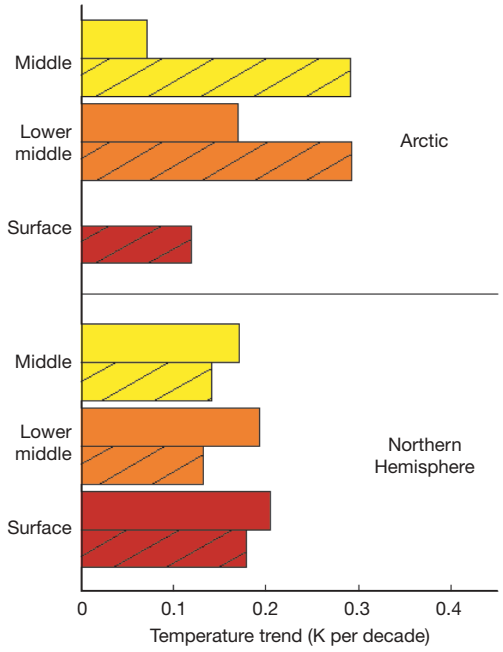


Figure 1 | Temperature trends over the Northern Hemisphere (0°–82.5° N) and the Arctic (65° N–82.5° N) for 1979–2001. Trends are for temperatures at the surface and in the lower-middle and middle troposphere from the ERA-40 reanalysis (hatched) and observations (solid) in the winter (December–February) season. The observed trends are derived from the HadCrut3v (ref. 8) data set for the surface temperature and from a satellite microwave sounding unit[9] (MSU; RSS version 3) for the temperatures in the lower-middle[10] and middle[11,12] troposphere. Observed surface temperatures in the Arctic are not shown, because they are spatially incomplete. For a direct comparison with the MSU observations, synthetic temperatures in the lower-middle and middle troposphere are computed from the ERA-40 reanalysis by applying the MSU weighting functions[3].

Hemisphere atmospheric warming in every season. Specifically, for trends in annual means in the reanalysis for 1979–2001, the Arctic warms 2.7 times more than the Northern Hemisphere in the lower-middle troposphere, in comparison with just 1.5 times more in the observations.

During the polar night, solar absorption at the surface is absent or weak. At the same time, the atmosphere transports a substantial amount of heat northwards from lower latitudes, with heating rates in the Arctic that maximize at about 1,500 m in winter[5]. For these reasons and others, strong radiative cooling at the surface causes frequent lower-tropospheric temperature inversions, which are very stable and damp vertical heat transfer during the polar night. When ice and snow retreat, some of the heat from increased solar absorption is stored at the ocean surface and is released during the cold seasons without warming the atmosphere aloft very much.

It has been concluded that northwards atmospheric heat transport into the Arctic should increase in a warming world[6,7] owing to increased evaporation in the tropics and subsequent condensation in the high latitudes. This increase in latent heat transport is somewhat counterbalanced by a decrease in sensible heat transport, as Arctic amplification decreases the pole-to-equator temperature gradient. Models indicate that warming aloft would not outpace the surface warming after considering increased northwards atmospheric heat transport along with the retreat of ice and snow[7]. Graversen *et al.*[2] find that the change in northwards atmospheric heat transport is not a substantial source of heating aloft in midwinter (January–February) in the Arctic.

The smaller warming trends aloft in the observations in winter are more consistent with the amplification of surface warming from ice and snow retreat and the lack of change in the northwards atmospheric heat transport for 1979–2001. This consistent set of observations calls into question the results of Graversen *et al.*[2] obtained for the polar night.

Cecilia M. Bitz[1] & Qiang Fu[1]

[1]Atmospheric Science Department, 408 Atmospheric Sciences and Geophysics Hall, University of Washington, Seattle, Washington 98195, USA.
e-mail: bitz@atmos.washington.edu

Received 5 February; accepted 9 May 2008.

1. Lemke, P. *et al.* in *Climate Change 2007: The Physical Science Basis* (eds Solomon, S. *et al.*) 337–383 (Contribution of Working Group I to the Fourth Assessment Report of the IPCC, Cambridge Univ. Press, 2007).
2. Graversen, R. G., Mauritsen, T., Tjernström, M., Källén, E. & Svensson, G. Vertical structure of recent Arctic warming. *Nature* **451,** 53–56 (2008).
3. Johanson, C. M. & Fu, Q. Antarctic atmospheric temperature trend patterns from satellite observations. *Geophys. Res. Lett.* **34,** doi:10.1029/2006GL029108 (2007).
4. Karl, T. R., Hassol, S. J., Miller, C. D. & Murray, W. L. (eds) *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences* (Synthesis and Assessment Product 1.1, US Climate Change Science Program, 2006).
5. Overland, J. E. & Turret, P. in *The Polar Oceans and Their Role in Shaping the Global Environment* (eds Johannessen, O. M., Muench, R. & Overland, J. E.) 313–325 (American Geophysical Union, 1994).
6. Alexseev, V. A., Langen, P. L. & Bates, J. R. Polar amplification of surface warming on an aquaplanet in "ghost forcing" experiments without sea ice feedbacks. *Clim. Dyn.* **24,** 655–666 (2005).
7. Cai, M. & Lu, J. Dynamical greenhouse-plus feedback and polar warming amplification. Part II: meridional and vertical asymmetries of the global warming. *Clim. Dyn.* **29,** 375–391 (2007).
8. Brohan, P., Kennedy, J. J., Haris, I., Tett, S. F. B. & Jones, P. D. Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *J. Geophys. Res.* **111,** doi:10.1029/2005JD006548 (2006).
9. Mears, C. A., Schabel, M. C. & Wentz, F. J. A reanalysis of the MSU channel 2 tropospheric temperature record. *J. Clim.* **16,** 3650–3664 (2003).
10. Mears, C. A. & Wentz, F. J. The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science* **309,** 1548–1551 (2005).
11. Fu, Q., Johanson, C. M., Warren, S. G. & Seidel, D. J. Contribution of stratospheric cooling to satellite-inferred tropospheric temperature trends. *Nature* **429,** 55–58 (2004).
12. Johanson, C. M. & Fu, Q. Robustness of tropospheric temperature trends from MSU channels 2 and 4. *J. Clim.* **19,** 4234–4242 (2006).

doi:10.1038/nature07258

Graversen et al. *reply*

Replying to: P. W. Thorne *Nature* 455, doi:10.1038/nature07256; A. N. Grant, S. Brönnimann & L. Haimberger *Nature* 455, doi:10.1038/nature07257; C. M. Bitz & Q. Fu *Nature* 455, doi:10.1038/nature07258 (2008)

These three communications[1–3] question the validity of some of our conclusions[4]. We found Arctic temperature trend amplification well above the boundary layer. In summer, the maximum amplification is found at a height of around 2 km, and no amplification is encountered near the surface. These findings appear in two state-of-the-art reanalyses, ERA-40 (ref. 5) and JRA-25 (ref. 6). Both these data sets show roughly the same overall vertical structure, and we believe our conclusions can be based on either of them. However, they show

considerable differences regarding the magnitudes of the Arctic trends (see our Supplementary Information[4]), but our conclusions are not based on the absolute magnitudes.

A reanalysis synthesizes all available observations and uses a physically based model of the atmosphere to weigh the observations against each other and to extrapolate the observed information in space and time to unobserved parts of the atmosphere. The assimilation procedure takes observational as well as model uncertainties into account. In ERA-40 and JRA-25, the strongest observational constraint on the Arctic temperatures aloft is provided by assimilation of satellite observations, such as microwave sounding unit (MSU) radiances, as *in situ* observations are few in this region. In the assimilation process, careful bias adjustment has been applied to the satellite observations[5].

We examined the agreement of the MSU satellite observations (RSS analysis[7] TLT v3.1 and TMT v3.2) with the vertical structure of ERA-40 and JRA-25. Arctic amplification is encountered in the channel representing the lower troposphere. In summer, both the lower-troposphere and the middle-troposphere channels indicate considerable warming over the Arctic. Because the Arctic surface temperature is constrained to be close to the melting point during this season, this warming must occur aloft, in accordance with the two reanalyses.

The annual lower-troposphere MSU trend reaches 0.46 K per decade at 81.25° N, calculated on the basis of a least-squares fit. We therefore find it surprising that Thorne[1] estimates a high-latitude trend of only 0.2 K per decade. Bitz and Fu[3] report winter trends in the lower troposphere of around 0.2 K per decade both for the Arctic and the Northern Hemisphere, which we also find. However, they report middle-troposphere trends of around 0.09 and 0.18 K per decade for the Arctic and the Northern Hemisphere, respectively. We find, on the other hand, 0.14 and 0.09 K per decade for the Arctic and the Northern Hemisphere, respectively. Hence, the MSU data show winter Arctic amplification in agreement with ERA-40 and JRA-25.

In their last paragraph, Bitz and Fu[3] indicate that ERA-40 exaggerates winter trends aloft. This might be the case; JRA-25 shows considerably smaller trends. However, our point is that, even in JRA-25, winter trends above the boundary layer are comparable to those near the surface and can hardly be linked to surface processes alone. Grant *et al.*[2] compare ERA-40 data with radiosonde observations, which are few in the Arctic. Although these observations cannot confirm the April–October warming aloft found in ERA-40, in general they show good agreement with the ERA-40 data at the points where radiosondes are available.

There is no doubt that more *in situ* observations in the Arctic are needed to enhance the quality of future reanalyses. Given the absence of such observations in historical archives, we feel that a reanalysis is likely to provide a better representation of the true state of the Arctic atmosphere than any single inhomogeneous set of a specific observation type. Satellite observations must be bias corrected and radio soundings exist almost only in the southern part of the Arctic. In a reanalysis, both of these shortcomings are consistently handled in the framework of a dynamical, global model of the atmosphere. We have given an estimate of the uncertainty associated with reanalysis data by displaying results from two different, second-generation reanalyses. Within the limits of this uncertainty we believe that our conclusions remain valid.

R. G. Graversen[1], T. Mauritsen[1], M. Tjernström[1], E. Källén[1] & G. Svensson[1]

[1]Department of Meteorology, Stockholm University, S-106 91 Stockholm, Sweden.

e-mail: rune@misu.su.se

1. Thorne, P. W. Arctic tropospheric warming amplification? *Nature* **455,** doi:10.1038/nature07256 (2008).
2. Grant, A. N., Brönnimann, S. & Haimberger, L. Recent Arctic warming vertical structure contested. *Nature* **455,** doi:10.1038/nature07257 (2008).
3. Bitz, C. M. & Fu, Q. Arctic warming aloft is data set dependent. *Nature* **455,** doi:10.1038/nature07258 (2008).
4. Graversen, R. G., Mauritsen, T., Tjernström, M., Källén, E. & Svensson, G. Vertical structure of recent Arctic warming. *Nature* **451,** 53–56 (2008).
5. Uppala, S. M. *et al.* The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.* **131,** 2961–3012 (2005).
6. Onogi, K. *et al.* The JRA-25 reanalysis. *J. Meteorol. Soc. Jpn* **85,** 369–432 (2007).
7. Mears, C. A. & Wentz, F. J. The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science* **309,** 1548–1551 (2005).

doi:10.1038/nature07259

Arctic tropospheric warming amplification?

Arising from: R. G. Graversen, T. Mauritsen, M. Tjernström, E. Källén & G. Svensson *Nature* **451**, 53–56 (2008)

Relative rates of temperature change between the troposphere and surface, and the mechanisms that produce these changes, have long been a contentious issue. Graversen *et al.*[1], predicated upon the ERA-40 reanalysis[2], report polar tropospheric amplification of surface warming and attempt to explain this finding dynamically. Here we show (1) that data from satellites[3,4] and weather balloons[5] indicate that the ERA-40 trends are increasingly unrealistic polewards of 62° N; (2) that the two other reanalyses considered[1] exhibit very different polar trends; and (3) that the vertical profile of polar trends in ERA-40 is unrealistic, particularly above the troposphere. These quasi-independent strands of evidence imply that the pattern of warming in the Arctic troposphere is highly unlikely to be as given in ERA-40 and as reported by Graversen *et al.*[1].

Reanalyses are numerical weather-prediction systems run in hindcast mode considering all globally available observations[2]. Strenuous efforts are made to take account of both time-varying biases in the data and the impacts of the very substantially changing mix and coverage of observations. However, many aspects of the long-term behaviour of reanalyses remain unreliable[6,7] and their suitability for use in monitoring atmospheric temperature trends has been questioned by a recent expert panel[8].

Comparing ERA-40 with several observational[3–5] 'lower tropospheric' retrievals (corresponding most closely with the original analysis, peaking at about 725 hPa) over the 62.5° N to 82.5° N latitude range (Fig. 1, left-hand panels) yields good month-to-month

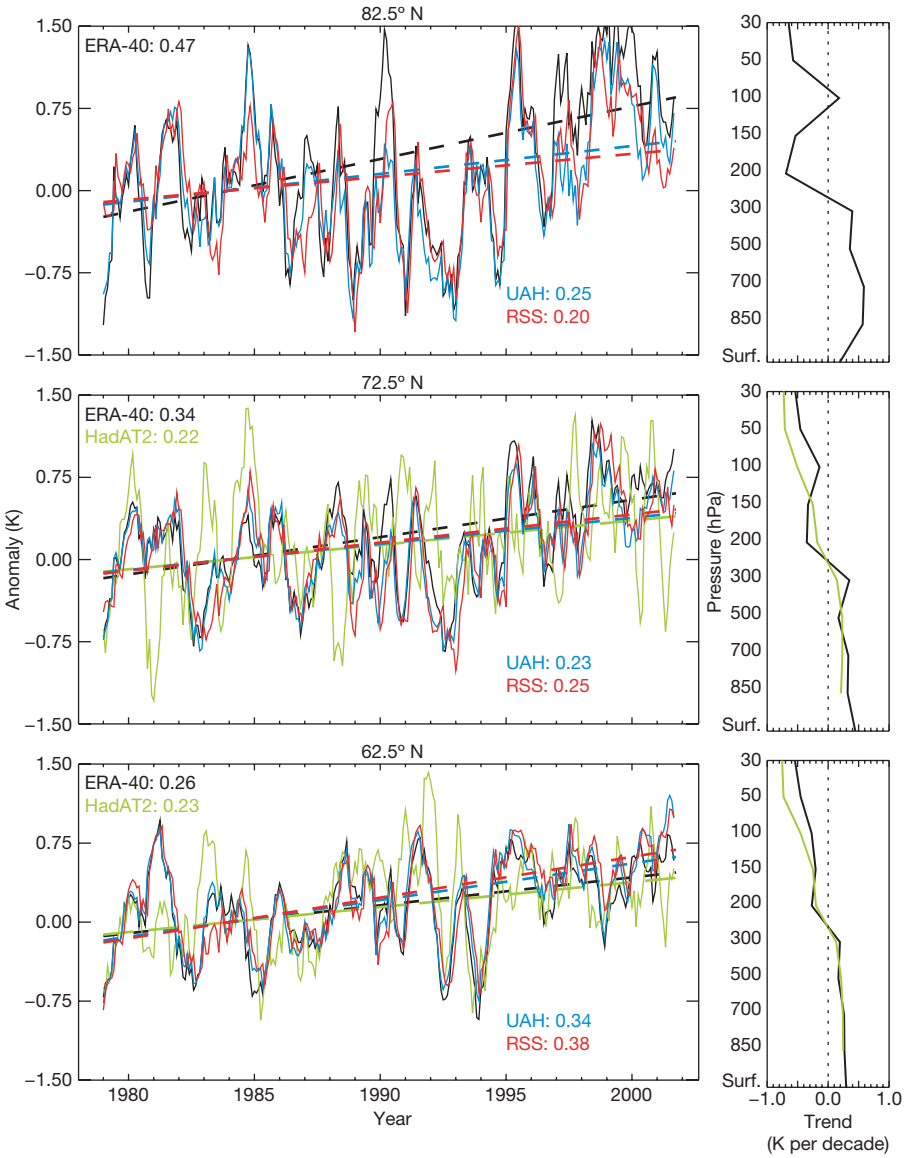


Figure 1 | Lower tropospheric retrieval data. Left-hand panels show temperature anomaly (relative to 1979–1988) monthly time series (smoothed with a simple seven-point moving filter) and trends (given as values in-line, for example ERA-40: 0.47) for three zonal bands for the broad T2LT lower tropospheric retrieval of the MSU record from UAH (ref. 3), RSS (ref. 4) and weighted equivalents from ERA-40 (ref. 2) and HadAT2 (ref. 5). Trends, calculated using a median-of-pairwise-slopes method[14], are quoted in kelvin per decade within each panel for the common period of record. Right-hand panels show vertically resolved trends on the nine HadAT2 levels for ERA-40 and HadAT2 (ref. 5). There are insufficient long-term radiosonde records at 82.5° N to assess climate trends, so there are no data here in HadAT2.

agreement, particularly with the globally complete satellite records, in accord with Graversen *et al.*[1]. Crucially, however, trends increasingly diverge as the pole is approached. High-frequency agreement is insufficient to ensure that the trend will be well characterized[9]. At 82.5° N, ERA-40 is overestimating the warming vis-à-vis available direct observational estimates by around 100%. It is north of about 80° N that ERA-40 shows the substantial warming reported by Graversen *et al.*[1]. At these latitudes, however, there are very few either conventional or space-based observations available to constrain the reanalyses. Therefore, the reality of these trends, given the lack of support from the available observational estimates[3,4] at 82.5° N, must be questioned.

Indeed, a comparison of Fig. 1 of Graversen *et al.*[1] with their Supplementary Figs 2 and 3 shows that the trend is not robust across different reanalyses systems. Although NCEP (ref. 10) can be considered a first-generation reanalysis, both ERA-40 (ref. 2) and the even newer JRA-25 (ref. 11) are second-generation reanalyses. The degree of pattern correspondence between these is visually poor, and the trend magnitudes differ substantially. This lack of robustness of the reported Arctic amplification signal implies that it is not necessarily a real-world feature.

Finally, a consideration of the full atmospheric profile rather than just that below 250 hPa shows that the ERA-40 trends become increasingly unrealistic with latitude (Fig. 1, right-hand panels). At 62.5° N, where radiosondes reporting temperatures, humidity and winds on distinct levels are plentiful, the ERA-40 trend looks realistic. Farther north, however, the availability of *in situ* radiosondes declines and the reanalysis is effectively unconstrained by *in situ* observations. Beyond 82.5° N, the reanalysis is constrained only by off-nadir views from infrared satellite observations. These are unlikely to be homogeneous. Furthermore, because they represent deep layers they cannot necessarily fully anchor the reanalysis temperatures, which may therefore have been affected by vertically differentiated model biases.

Taken together, the evidence implies that the reported Arctic tropospheric amplification is a non-climatic artefact in ERA-40. This reinforces the importance of treating any single data set, be it observational or derived, with extreme caution[12]. It does not imply that current reanalyses are unfit for the majority of purposes to which they are put. It does, however, reaffirm the importance of a properly resourced and scientifically robust attempt to create a truly climate-quality reanalysis product: a product that adequately retains long-term trend fidelity in all meteorological parameters[13].

Peter W. Thorne[1]
[1]Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK.
e-mail: peter.thorne@metoffice.gov.uk

Received 16 January; accepted 9 May 2008.

1. Graversen, R. G., Mauritsen, T., Tjernström, M., Källén, E. & Svensson, G. Vertical structure of recent Arctic warming. *Nature* **451,** 53–56 (2008).
2. Uppala, S. M. *et al.* The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.* **131,** 2961–3012 (2005).
3. Christy, J. R., Spencer, R. W., Norris, W. B., Braswell, W. D. & Parker, D. E. Error estimates of version 5.0 of MSU-AMSU bulk atmospheric temperatures. *J. Atmos. Ocean. Technol.* **20,** 613–629 (2003).
4. Mears, C. A. & Wentz, F. J. The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science* **309,** 1548–1551 (2005).
5. Thorne, P. W. *et al.* Revisiting radiosonde upper air temperatures from 1958 to 2002. *J. Geophys. Res.* **110,** D18105 (2005).
6. Mears, C. A., Santer, B. D., Wentz, F. J., Taylor, K. E. & Wehner, M. F. Relationship between temperature and precipitable water changes over tropical oceans. *Geophys. Res. Lett.* **34,** doi:10.1029/2007GL031936 (2007).
7. Bengtsson, L., Hodges, K. I. & Hagemann, S. Sensitivity of the ERA40 reanalysis to the observing system: determination of the global atmospheric circulation from reduced observations. *Tellus A* **56,** 456–471 (2004).
8. Karl, T. R., Hassol, S. J., Miller, C. D. & Murray, W. L. (eds) *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences* (Synthesis and Assessment Product 1.1, US Climate Change Science Program, 2006).
9. Sherwood, S. C., Titchner, H. A., Thorne, P. W. & McCarthy, M. C. How do we tell which estimates of past climate change are correct? *Int. J. Climatol.* (submitted).
10. Kalnay, E. *et al.* The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77,** 437–470 (1996).
11. Onogi, K. *et al.* The JRA-25 reanalysis. *J. Meteorol. Soc. Jpn* **85,** 369–432 (2007).
12. Thorne, P. W., Parker, D. E., Christy, J. R. & Mears, C. A. Uncertainties in climate trends - Lessons from upper-air temperature records. *Bull. Am. Meteorol. Soc.* **86,** 1437–1442 (2005).
13. Bengtsson, L. *et al.* The need for a dynamical climate reanalysis. *Bull. Am. Meteorol. Soc.* **88,** 495–501 (2007).
14. Lanzante, J. R. Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data. *Int. J. Climatol.* **16,** 1197–1226 (1996).

doi:10.1038/nature07256

Recent Arctic warming vertical structure contested

Arising from: R. G. Graversen, T. Mauritsen, M. Tjernström, E. Källén & G. Svensson *Nature* 451, 53–56 (2008)

The vertical structure of the recent Arctic warming contains information about the processes governing Arctic climate trends. Graversen *et al.* argue[1], on the basis of ERA-40 reanalysis[2] data, that a distinct maximum in 1979–2001 warm-season (April–October) Arctic temperature trends appears around 3 km above ground. Here we show that this is due to the heterogeneous nature of the data source, which incorporates information from satellites and radiosondes. Radiosonde data alone suggest the warming was strongest near ground.

Graversen *et al.*[1] claim that the warm-season temperature trend has a maximum at around 700 hPa, polewards of 75° N, and argue that anomalous heat advection from more southerly latitudes is important. However, the ERA-40 reanalysis may not be suitable for trend analysis as it incorporates information from different observing systems such as satellite and radiosonde, which might be inconsistent, in particular with respect to trends[3,4]. Radiosonde measurements provide vertically resolved temperature profiles in the troposphere, whereas satellites provide information on a weighted average over a thick layer. Furthermore, the ERA-40 assimilation system extrapolates information from data-rich to data-sparse areas, which is less reliable than observations. The ERA-40 reanalysis in the polar region has not been sufficiently validated by *in situ* observations and documented[2,5] problems with satellite radiance assimilations over the Arctic Ocean could lead to spurious trends.

A map of warm-season trends at 700 hPa (the peak level of the polar warming trend in ref. 1) from ERA-40 and radiosonde observations[6,7] confirms that the enhanced warming signal lies mostly in areas with no radiosonde data coverage (Fig. 1a). This is particularly so polewards of 75° N, where the trend appears strongest in ref. 1. Moreover, the few radiosonde data available near or polewards of 75° N show modest trends. To illustrate the effects on the vertical structure of the trend, we calculated zonally averaged vertical temperature trends from (1) ERA-40 reanalysis data, (2) such data subsampled to locations where radiosonde information is available (that is, where ERA-40 is best constrained) and (3) from only radiosonde data. The trend in the reanalysis (Fig. 1b) is a reproduction of Fig. 4a in ref. 1 and exhibits a maximum at 700 hPa, polewards of 75° N. Subsampling the ERA-40 reanalysis (Fig. 1c) reveals clearly different trends, and calculating trends directly from radiosondes alters matters even further (Fig. 1d). The result is independent of the methods used to homogenize the radiosonde data (unadjusted, RAOBCORE v1.4 (ref. 7) and RICH (ref. 8); not shown). The radiosonde data (note that some regions are not well covered and some levels are

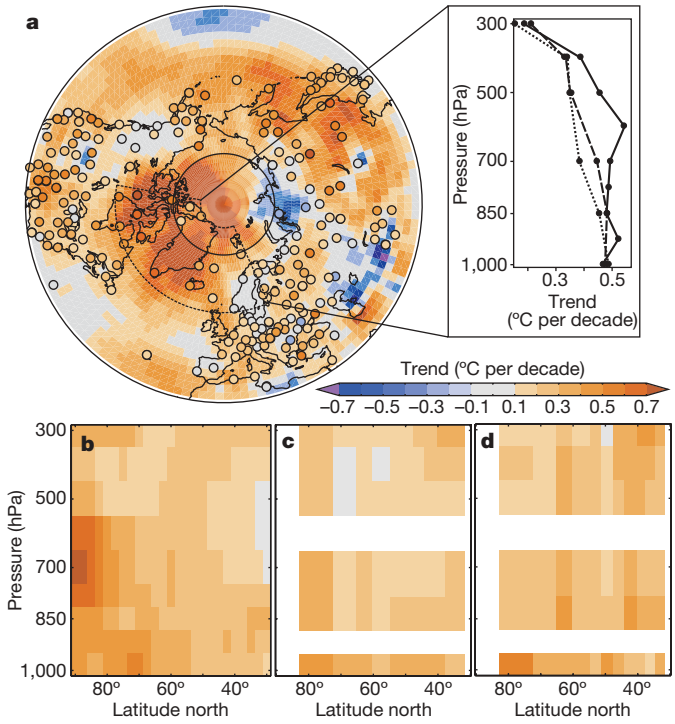


Figure 1 | Vertical structure of Arctic temperature trends for April to October, 1979–2001. Trends were calculated from seasonally averaged monthly anomalies using least-squares regression (not more than one missing month per season allowed, not more than five missing seasons in 1979–2001, neither first nor last two years can be missing). **a**, Trend field at 700 hPa from ERA-40 (ref. 2) and from radiosonde data[6,7] (circles) with 75° N latitude circle indicated by the thin solid line. **b**, Trends of zonal-mean temperature as a function of latitude and altitude from ERA-40. **c**, Same as **b**, but from ERA-40 subsampled to the locations and times where radiosonde data are available (anomalies zonally averaged in equal-area latitude bands). **d**, Same as **c**, but for radiosonde data. Inset in **a**, average trend profiles of the region 58° N–82° N, 100° W–25° E for full ERA-40 (solid line), subsampled ERA-40 (dashed) and radiosonde data (dotted).

missing because of inconsistent reporting) have their strongest trend near the ground, not above the boundary layer as in the full reanalysis. This is important because boundary layer processes are much more locally driven and simultaneously not well represented in a reanalysis. The same result is found when analysing a subregion with relatively even radiosonde coverage (inset, Fig. 1a), and during the remainder of the year (not shown).

Arctic climate is controlled by processes operating on scales from local to global, including transport effects; forcings such as greenhouse gases, aerosols and clouds; and feedbacks such as the well-known sea-ice–albedo feedback. The temperature profile can be a clue to the underlying processes, but to disentangle the contributions to Arctic temperature trends fully, vertical temperature structures should be addressed in a regionally and seasonally resolved manner. Furthermore, the large interannual variability in the Arctic, coupled with the sensitivity of trends to both end points and season definitions, suggests care should be taken in interpreting trends over short periods.

In conclusion, some features of the temperature trends calculated in ref. 1 reflect possible inhomogeneities or artefacts in the ERA-40 reanalysis rather than true climate signals, as they appear not to be supported by observations. ERA-40 reanalysis is a valuable tool in calculating circulation effects, especially on a subdecadal basis, but inhomogeneities and gaps in the global observing system tend to make trends from reanalyses unreliable, particularly in data-sparse regions.

A. N. Grant[1], **S. Brönnimann**[1] **& L. Haimberger**[2]
[1]Institute for Atmospheric and Climate Science, ETH Zurich, Universitätstrasse 16, CH-8092 Zurich, Switzerland.
e-mail: andrea.grant@env.ethz.ch
[2]Department of Meteorology and Geophysics, University of Vienna, Althanstrasse 14, A-1090 Vienna, Austria.

Received 23 January; accepted 12 May 2008.

1. Graversen, R. G., Mauritsen, T., Tjernström, M., Källén, E. & Svensson, G. Vertical structure of recent Arctic warming. *Nature* **451**, 53–56 (2008).
2. Uppala, S. M. *et al.* The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.* **131**, 2961–3012 (2005).
3. Fu, Q., Johanson, C. M., Warren, S. G. & Seidel, D. J. Contribution of stratospheric cooling to satellite-inferred tropospheric temperature trends. *Nature* **429**, 55–58 (2004).
4. Mears, C. A. & Wentz, F. J. The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science* **309**, 1548–1551 (2005).
5. Bromwich, D. & Wang, S.-H. Evaluation of the NCEP-NCAR and ECMWF 15- and 40-yr reanalyses using rawindsonde data from two independent Arctic field experiments. *Mon. Weath. Rev.* **133**, 3562–3578 (2005).
6. Durre, I., Vose, R. & Wuertz, D. Overview of the integrated global radiosonde archive. *J. Clim.* **19**, 53–68 (2006).
7. Haimberger, L. Homogenization of radiosonde temperature time series using innovation statistics. *J. Clim.* **20**, 1377–1403 (2007).
8. Haimberger, L. *et al.* Towards elimination of the warm bias in historic radiosonde temperature records — some new results from a comprehensive intercomparison of upper air data. *J. Clim.* (in the press).

doi:10.1038/nature07257

Arctic warming aloft is data set dependent

Arising from: R. G. Graversen, T. Mauritsen, M. Tjernström, E. Källén & G. Svensson *Nature* 451, 53–56 (2008)

Arctic sea ice and snow on land have retreated polewards at an alarming pace in the past few decades[1]. Such retreat locally amplifies surface warming through a positive feedback, which causes the Arctic surface to warm faster than the rest of the globe. In contrast, ice and snow retreat causes little warming in the atmosphere above when the stable winter atmosphere inhibits vertical heat exchange. We therefore find surprising the recent report by Graversen *et al.*[2] in which they claim that recent Arctic atmospheric warming extends far deeper into the atmosphere than expected, and can even exceed the surface warming during the polar night. Using a different data set, we show that there is much less warming aloft in winter, consistent with the recent retreat of ice and snow, as well as recent changes in atmospheric heat transport.

Graversen *et al.*[2] compute trends for 1979–2001 from ERA-40 reanalysis, which is a hybrid product using many types of raw observational data assimilated with a consistent global analysis system. The assimilation compensates for some but not all of the variations in the observing system over time that may compromise the veracity of the temperature trend analyses[3]. Figure 1 compares temperature trends in winter from the ERA-40 reanalysis with climate-quality records from satellite observations[4]. Trends in the Arctic winter aloft are strongly data set dependent: the observed trend is 75% less than reanalysis in the middle troposphere and 40% less than in the lower-middle troposphere. In comparison with the observations, the reanalysis exaggerates polar amplification aloft by overestimating the Arctic atmospheric warming and underestimating the Northern

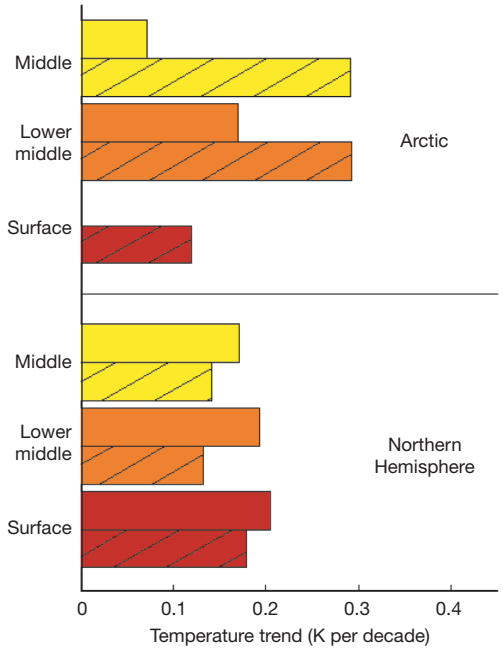


Figure 1 | Temperature trends over the Northern Hemisphere (0°–82.5° N) and the Arctic (65° N–82.5° N) for 1979–2001. Trends are for temperatures at the surface and in the lower-middle and middle troposphere from the ERA-40 reanalysis (hatched) and observations (solid) in the winter (December–February) season. The observed trends are derived from the HadCrut3v (ref. 8) data set for the surface temperature and from a satellite microwave sounding unit[9] (MSU; RSS version 3) for the temperatures in the lower-middle[10] and middle[11,12] troposphere. Observed surface temperatures in the Arctic are not shown, because they are spatially incomplete. For a direct comparison with the MSU observations, synthetic temperatures in the lower-middle and middle troposphere are computed from the ERA-40 reanalysis by applying the MSU weighting functions[3].

Hemisphere atmospheric warming in every season. Specifically, for trends in annual means in the reanalysis for 1979–2001, the Arctic warms 2.7 times more than the Northern Hemisphere in the lower-middle troposphere, in comparison with just 1.5 times more in the observations.

During the polar night, solar absorption at the surface is absent or weak. At the same time, the atmosphere transports a substantial amount of heat northwards from lower latitudes, with heating rates in the Arctic that maximize at about 1,500 m in winter[5]. For these reasons and others, strong radiative cooling at the surface causes frequent lower-tropospheric temperature inversions, which are very stable and damp vertical heat transfer during the polar night. When ice and snow retreat, some of the heat from increased solar absorption is stored at the ocean surface and is released during the cold seasons without warming the atmosphere aloft very much.

It has been concluded that northwards atmospheric heat transport into the Arctic should increase in a warming world[6,7] owing to increased evaporation in the tropics and subsequent condensation in the high latitudes. This increase in latent heat transport is somewhat counterbalanced by a decrease in sensible heat transport, as Arctic amplification decreases the pole-to-equator temperature gradient. Models indicate that warming aloft would not outpace the surface warming after considering increased northwards atmospheric heat transport along with the retreat of ice and snow[7]. Graversen *et al.*[2] find that the change in northwards atmospheric heat transport is not a substantial source of heating aloft in midwinter (January–February) in the Arctic.

The smaller warming trends aloft in the observations in winter are more consistent with the amplification of surface warming from ice and snow retreat and the lack of change in the northwards atmospheric heat transport for 1979–2001. This consistent set of observations calls into question the results of Graversen *et al.*[2] obtained for the polar night.

Cecilia M. Bitz[1] & Qiang Fu[1]

[1]Atmospheric Science Department, 408 Atmospheric Sciences and Geophysics Hall, University of Washington, Seattle, Washington 98195, USA.
e-mail: bitz@atmos.washington.edu

Received 5 February; accepted 9 May 2008.

1. Lemke, P. *et al.* in *Climate Change 2007: The Physical Science Basis* (eds Solomon, S. *et al.*) 337–383 (Contribution of Working Group I to the Fourth Assessment Report of the IPCC, Cambridge Univ. Press, 2007).
2. Graversen, R. G., Mauritsen, T., Tjernström, M., Källén, E. & Svensson, G. Vertical structure of recent Arctic warming. *Nature* **451,** 53–56 (2008).
3. Johanson, C. M. & Fu, Q. Antarctic atmospheric temperature trend patterns from satellite observations. *Geophys. Res. Lett.* **34,** doi:10.1029/2006GL029108 (2007).
4. Karl, T. R., Hassol, S. J., Miller, C. D. & Murray, W. L. (eds) *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences* (Synthesis and Assessment Product 1.1, US Climate Change Science Program, 2006).
5. Overland, J. E. & Turret, P. in *The Polar Oceans and Their Role in Shaping the Global Environment* (eds Johannessen, O. M., Muench, R. & Overland, J. E.) 313–325 (American Geophysical Union, 1994).
6. Alexseev, V. A., Langen, P. L. & Bates, J. R. Polar amplification of surface warming on an aquaplanet in "ghost forcing" experiments without sea ice feedbacks. *Clim. Dyn.* **24,** 655–666 (2005).
7. Cai, M. & Lu, J. Dynamical greenhouse-plus feedback and polar warming amplification. Part II: meridional and vertical asymmetries of the global warming. *Clim. Dyn.* **29,** 375–391 (2007).
8. Brohan, P., Kennedy, J. J., Haris, I., Tett, S. F. B. & Jones, P. D. Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *J. Geophys. Res.* **111,** doi:10.1029/2005JD006548 (2006).
9. Mears, C. A., Schabel, M. C. & Wentz, F. J. A reanalysis of the MSU channel 2 tropospheric temperature record. *J. Clim.* **16,** 3650–3664 (2003).
10. Mears, C. A. & Wentz, F. J. The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science* **309,** 1548–1551 (2005).
11. Fu, Q., Johanson, C. M., Warren, S. G. & Seidel, D. J. Contribution of stratospheric cooling to satellite-inferred tropospheric temperature trends. *Nature* **429,** 55–58 (2004).
12. Johanson, C. M. & Fu, Q. Robustness of tropospheric temperature trends from MSU channels 2 and 4. *J. Clim.* **19,** 4234–4242 (2006).

doi:10.1038/nature07258

Graversen et al. *reply*

Replying to: P. W. Thorne *Nature* 455, doi:10.1038/nature07256; A. N. Grant, S. Brönnimann & L. Haimberger *Nature* 455, doi:10.1038/nature07257; C. M. Bitz & Q. Fu *Nature* 455, doi:10.1038/nature07258 (2008)

These three communications[1–3] question the validity of some of our conclusions[4]. We found Arctic temperature trend amplification well above the boundary layer. In summer, the maximum amplification is found at a height of around 2 km, and no amplification is encountered near the surface. These findings appear in two state-of-the-art reanalyses, ERA-40 (ref. 5) and JRA-25 (ref. 6). Both these data sets show roughly the same overall vertical structure, and we believe our conclusions can be based on either of them. However, they show

considerable differences regarding the magnitudes of the Arctic trends (see our Supplementary Information[4]), but our conclusions are not based on the absolute magnitudes.

A reanalysis synthesizes all available observations and uses a physically based model of the atmosphere to weigh the observations against each other and to extrapolate the observed information in space and time to unobserved parts of the atmosphere. The assimilation procedure takes observational as well as model uncertainties into account. In ERA-40 and JRA-25, the strongest observational constraint on the Arctic temperatures aloft is provided by assimilation of satellite observations, such as microwave sounding unit (MSU) radiances, as *in situ* observations are few in this region. In the assimilation process, careful bias adjustment has been applied to the satellite observations[5].

We examined the agreement of the MSU satellite observations (RSS analysis[7] TLT v3.1 and TMT v3.2) with the vertical structure of ERA-40 and JRA-25. Arctic amplification is encountered in the channel representing the lower troposphere. In summer, both the lower-troposphere and the middle-troposphere channels indicate considerable warming over the Arctic. Because the Arctic surface temperature is constrained to be close to the melting point during this season, this warming must occur aloft, in accordance with the two reanalyses.

The annual lower-troposphere MSU trend reaches 0.46 K per decade at 81.25° N, calculated on the basis of a least-squares fit. We therefore find it surprising that Thorne[1] estimates a high-latitude trend of only 0.2 K per decade. Bitz and Fu[3] report winter trends in the lower troposphere of around 0.2 K per decade both for the Arctic and the Northern Hemisphere, which we also find. However, they report middle-troposphere trends of around 0.09 and 0.18 K per decade for the Arctic and the Northern Hemisphere, respectively. We find, on the other hand, 0.14 and 0.09 K per decade for the Arctic and the Northern Hemisphere, respectively. Hence, the MSU data show winter Arctic amplification in agreement with ERA-40 and JRA-25.

In their last paragraph, Bitz and Fu[3] indicate that ERA-40 exaggerates winter trends aloft. This might be the case; JRA-25 shows considerably smaller trends. However, our point is that, even in JRA-25, winter trends above the boundary layer are comparable to those near the surface and can hardly be linked to surface processes alone. Grant *et al.*[2] compare ERA-40 data with radiosonde observations, which are few in the Arctic. Although these observations cannot confirm the April–October warming aloft found in ERA-40, in general they show good agreement with the ERA-40 data at the points where radiosondes are available.

There is no doubt that more *in situ* observations in the Arctic are needed to enhance the quality of future reanalyses. Given the absence of such observations in historical archives, we feel that a reanalysis is likely to provide a better representation of the true state of the Arctic atmosphere than any single inhomogeneous set of a specific observation type. Satellite observations must be bias corrected and radio soundings exist almost only in the southern part of the Arctic. In a reanalysis, both of these shortcomings are consistently handled in the framework of a dynamical, global model of the atmosphere. We have given an estimate of the uncertainty associated with reanalysis data by displaying results from two different, second-generation reanalyses. Within the limits of this uncertainty we believe that our conclusions remain valid.

R. G. Graversen[1], T. Mauritsen[1], M. Tjernström[1], E. Källén[1] & G. Svensson[1]

[1]Department of Meteorology, Stockholm University, S-106 91 Stockholm, Sweden.

e-mail: rune@misu.su.se

1. Thorne, P. W. Arctic tropospheric warming amplification? *Nature* **455,** doi:10.1038/nature07256 (2008).
2. Grant, A. N., Brönnimann, S. & Haimberger, L. Recent Arctic warming vertical structure contested. *Nature* **455,** doi:10.1038/nature07257 (2008).
3. Bitz, C. M. & Fu, Q. Arctic warming aloft is data set dependent. *Nature* **455,** doi:10.1038/nature07258 (2008).
4. Graversen, R. G., Mauritsen, T., Tjernström, M., Källén, E. & Svensson, G. Vertical structure of recent Arctic warming. *Nature* **451,** 53–56 (2008).
5. Uppala, S. M. *et al.* The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.* **131,** 2961–3012 (2005).
6. Onogi, K. *et al.* The JRA-25 reanalysis. *J. Meteorol. Soc. Jpn* **85,** 369–432 (2007).
7. Mears, C. A. & Wentz, F. J. The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science* **309,** 1548–1551 (2005).

doi:10.1038/nature07259

Arctic tropospheric warming amplification?

Arising from: R. G. Graversen, T. Mauritsen, M. Tjernström, E. Källén & G. Svensson *Nature* **451**, 53–56 (2008)

Relative rates of temperature change between the troposphere and surface, and the mechanisms that produce these changes, have long been a contentious issue. Graversen *et al.*[1], predicated upon the ERA-40 reanalysis[2], report polar tropospheric amplification of surface warming and attempt to explain this finding dynamically. Here we show (1) that data from satellites[3,4] and weather balloons[5] indicate that the ERA-40 trends are increasingly unrealistic polewards of 62° N; (2) that the two other reanalyses considered[1] exhibit very different polar trends; and (3) that the vertical profile of polar trends in ERA-40 is unrealistic, particularly above the troposphere. These quasi-independent strands of evidence imply that the pattern of warming in the Arctic troposphere is highly unlikely to be as given in ERA-40 and as reported by Graversen *et al.*[1].

Reanalyses are numerical weather-prediction systems run in hindcast mode considering all globally available observations[2]. Strenuous efforts are made to take account of both time-varying biases in the data and the impacts of the very substantially changing mix and coverage of observations. However, many aspects of the long-term behaviour of reanalyses remain unreliable[6,7] and their suitability for use in monitoring atmospheric temperature trends has been questioned by a recent expert panel[8].

Comparing ERA-40 with several observational[3–5] 'lower tropospheric' retrievals (corresponding most closely with the original analysis, peaking at about 725 hPa) over the 62.5° N to 82.5° N latitude range (Fig. 1, left-hand panels) yields good month-to-month

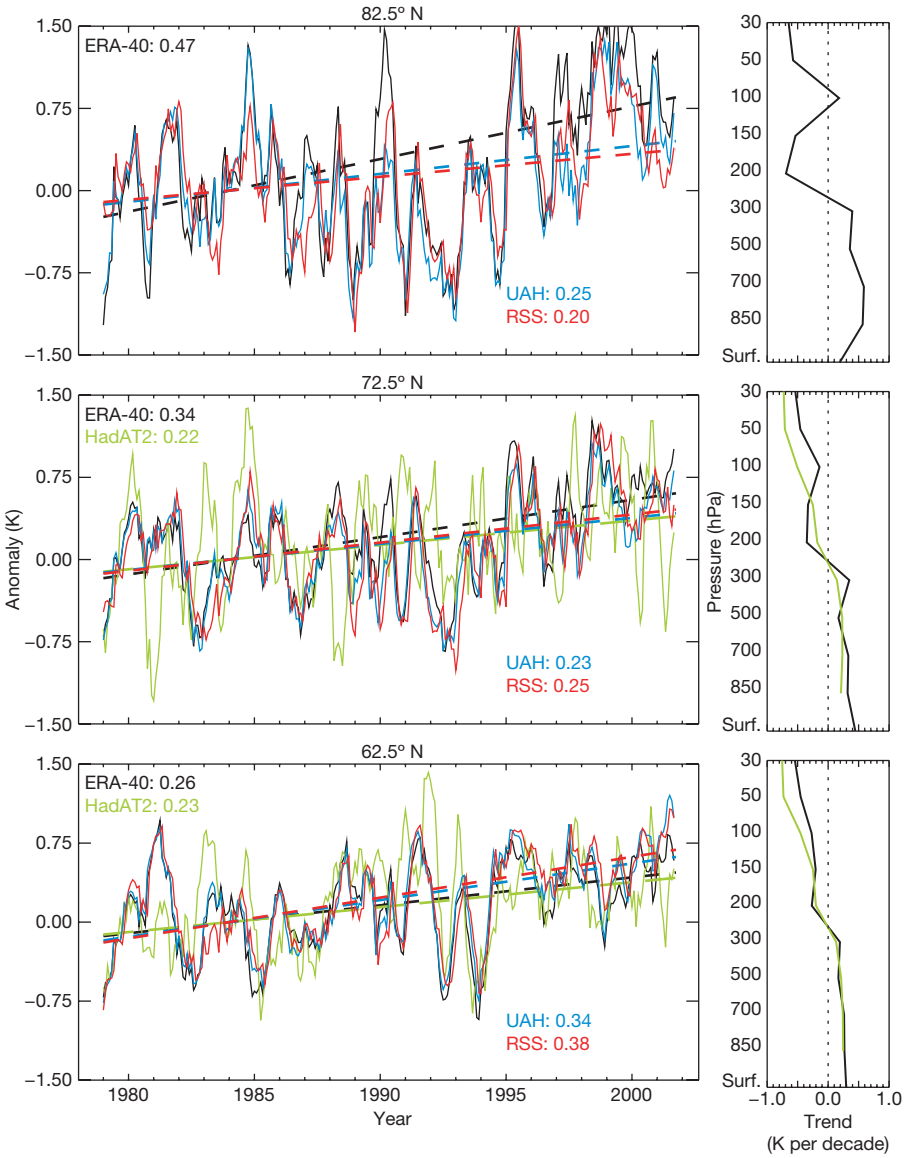


Figure 1 | Lower tropospheric retrieval data. Left-hand panels show temperature anomaly (relative to 1979–1988) monthly time series (smoothed with a simple seven-point moving filter) and trends (given as values in-line, for example ERA-40: 0.47) for three zonal bands for the broad T2LT lower tropospheric retrieval of the MSU record from UAH (ref. 3), RSS (ref. 4) and weighted equivalents from ERA-40 (ref. 2) and HadAT2 (ref. 5). Trends, calculated using a median-of-pairwise-slopes method[14], are quoted in kelvin per decade within each panel for the common period of record. Right-hand panels show vertically resolved trends on the nine HadAT2 levels for ERA-40 and HadAT2 (ref. 5). There are insufficient long-term radiosonde records at 82.5° N to assess climate trends, so there are no data here in HadAT2.

agreement, particularly with the globally complete satellite records, in accord with Graversen *et al.*[1]. Crucially, however, trends increasingly diverge as the pole is approached. High-frequency agreement is insufficient to ensure that the trend will be well characterized[9]. At 82.5° N, ERA-40 is overestimating the warming vis-à-vis available direct observational estimates by around 100%. It is north of about 80° N that ERA-40 shows the substantial warming reported by Graversen *et al.*[1]. At these latitudes, however, there are very few either conventional or space-based observations available to constrain the reanalyses. Therefore, the reality of these trends, given the lack of support from the available observational estimates[3,4] at 82.5° N, must be questioned.

Indeed, a comparison of Fig. 1 of Graversen *et al.*[1] with their Supplementary Figs 2 and 3 shows that the trend is not robust across different reanalyses systems. Although NCEP (ref. 10) can be considered a first-generation reanalysis, both ERA-40 (ref. 2) and the even newer JRA-25 (ref. 11) are second-generation reanalyses. The degree of pattern correspondence between these is visually poor, and the trend magnitudes differ substantially. This lack of robustness of the reported Arctic amplification signal implies that it is not necessarily a real-world feature.

Finally, a consideration of the full atmospheric profile rather than just that below 250 hPa shows that the ERA-40 trends become increasingly unrealistic with latitude (Fig. 1, right-hand panels). At 62.5° N, where radiosondes reporting temperatures, humidity and winds on distinct levels are plentiful, the ERA-40 trend looks realistic. Farther north, however, the availability of *in situ* radiosondes declines and the reanalysis is effectively unconstrained by *in situ* observations. Beyond 82.5° N, the reanalysis is constrained only by off-nadir views from infrared satellite observations. These are unlikely to be homogeneous. Furthermore, because they represent deep layers they cannot necessarily fully anchor the reanalysis temperatures, which may therefore have been affected by vertically differentiated model biases.

Taken together, the evidence implies that the reported Arctic tropospheric amplification is a non-climatic artefact in ERA-40. This reinforces the importance of treating any single data set, be it observational or derived, with extreme caution[12]. It does not imply that current reanalyses are unfit for the majority of purposes to which they are put. It does, however, reaffirm the importance of a properly resourced and scientifically robust attempt to create a truly climate-quality reanalysis product: a product that adequately retains long-term trend fidelity in all meteorological parameters[13].

Peter W. Thorne[1]
[1]Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK.
e-mail: peter.thorne@metoffice.gov.uk

Received 16 January; accepted 9 May 2008.

1. Graversen, R. G., Mauritsen, T., Tjernström, M., Källén, E. & Svensson, G. Vertical structure of recent Arctic warming. *Nature* **451,** 53–56 (2008).
2. Uppala, S. M. *et al.* The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.* **131,** 2961–3012 (2005).
3. Christy, J. R., Spencer, R. W., Norris, W. B., Braswell, W. D. & Parker, D. E. Error estimates of version 5.0 of MSU-AMSU bulk atmospheric temperatures. *J. Atmos. Ocean. Technol.* **20,** 613–629 (2003).
4. Mears, C. A. & Wentz, F. J. The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science* **309,** 1548–1551 (2005).
5. Thorne, P. W. *et al.* Revisiting radiosonde upper air temperatures from 1958 to 2002. *J. Geophys. Res.* **110,** D18105 (2005).
6. Mears, C. A., Santer, B. D., Wentz, F. J., Taylor, K. E. & Wehner, M. F. Relationship between temperature and precipitable water changes over tropical oceans. *Geophys. Res. Lett.* **34,** doi:10.1029/2007GL031936 (2007).
7. Bengtsson, L., Hodges, K. I. & Hagemann, S. Sensitivity of the ERA40 reanalysis to the observing system: determination of the global atmospheric circulation from reduced observations. *Tellus A* **56,** 456–471 (2004).
8. Karl, T. R., Hassol, S. J., Miller, C. D. & Murray, W. L. (eds) *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences* (Synthesis and Assessment Product 1.1, US Climate Change Science Program, 2006).
9. Sherwood, S. C., Titchner, H. A., Thorne, P. W. & McCarthy, M. C. How do we tell which estimates of past climate change are correct? *Int. J. Climatol.* (submitted).
10. Kalnay, E. *et al.* The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77,** 437–470 (1996).
11. Onogi, K. *et al.* The JRA-25 reanalysis. *J. Meteorol. Soc. Jpn* **85,** 369–432 (2007).
12. Thorne, P. W., Parker, D. E., Christy, J. R. & Mears, C. A. Uncertainties in climate trends - Lessons from upper-air temperature records. *Bull. Am. Meteorol. Soc.* **86,** 1437–1442 (2005).
13. Bengtsson, L. *et al.* The need for a dynamical climate reanalysis. *Bull. Am. Meteorol. Soc.* **88,** 495–501 (2007).
14. Lanzante, J. R. Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data. *Int. J. Climatol.* **16,** 1197–1226 (1996).

doi:10.1038/nature07256

Recent Arctic warming vertical structure contested

Arising from: R. G. Graversen, T. Mauritsen, M. Tjernström, E. Källén & G. Svensson *Nature* 451, 53–56 (2008)

The vertical structure of the recent Arctic warming contains information about the processes governing Arctic climate trends. Graversen *et al.* argue[1], on the basis of ERA-40 reanalysis[2] data, that a distinct maximum in 1979–2001 warm-season (April–October) Arctic temperature trends appears around 3 km above ground. Here we show that this is due to the heterogeneous nature of the data source, which incorporates information from satellites and radiosondes. Radiosonde data alone suggest the warming was strongest near ground.

Graversen *et al.*[1] claim that the warm-season temperature trend has a maximum at around 700 hPa, polewards of 75° N, and argue that anomalous heat advection from more southerly latitudes is important. However, the ERA-40 reanalysis may not be suitable for trend analysis as it incorporates information from different observing systems such as satellite and radiosonde, which might be inconsistent, in particular with respect to trends[3,4]. Radiosonde measurements provide vertically resolved temperature profiles in the troposphere, whereas satellites provide information on a weighted average over a thick layer. Furthermore, the ERA-40 assimilation system extrapolates information from data-rich to data-sparse areas, which is less reliable than observations. The ERA-40 reanalysis in the polar region has not been sufficiently validated by *in situ* observations and documented[2,5] problems with satellite radiance assimilations over the Arctic Ocean could lead to spurious trends.

A map of warm-season trends at 700 hPa (the peak level of the polar warming trend in ref. 1) from ERA-40 and radiosonde observations[6,7] confirms that the enhanced warming signal lies mostly in areas with no radiosonde data coverage (Fig. 1a). This is particularly so polewards of 75° N, where the trend appears strongest in ref. 1. Moreover, the few radiosonde data available near or polewards of 75° N show modest trends. To illustrate the effects on the vertical structure of the trend, we calculated zonally averaged vertical temperature trends from (1) ERA-40 reanalysis data, (2) such data subsampled to locations where radiosonde information is available (that is, where ERA-40 is best constrained) and (3) from only radiosonde data. The trend in the reanalysis (Fig. 1b) is a reproduction of Fig. 4a in ref. 1 and exhibits a maximum at 700 hPa, polewards of 75° N. Subsampling the ERA-40 reanalysis (Fig. 1c) reveals clearly different trends, and calculating trends directly from radiosondes alters matters even further (Fig. 1d). The result is independent of the methods used to homogenize the radiosonde data (unadjusted, RAOBCORE v1.4 (ref. 7) and RICH (ref. 8); not shown). The radiosonde data (note that some regions are not well covered and some levels are

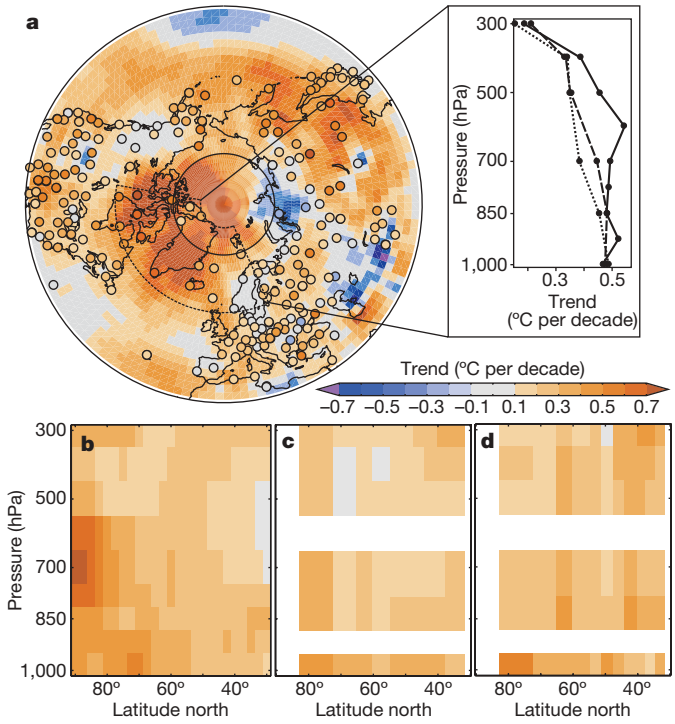


Figure 1 | Vertical structure of Arctic temperature trends for April to October, 1979–2001. Trends were calculated from seasonally averaged monthly anomalies using least-squares regression (not more than one missing month per season allowed, not more than five missing seasons in 1979–2001, neither first nor last two years can be missing). **a**, Trend field at 700 hPa from ERA-40 (ref. 2) and from radiosonde data[6,7] (circles) with 75° N latitude circle indicated by the thin solid line. **b**, Trends of zonal-mean temperature as a function of latitude and altitude from ERA-40. **c**, Same as **b**, but from ERA-40 subsampled to the locations and times where radiosonde data are available (anomalies zonally averaged in equal-area latitude bands). **d**, Same as **c**, but for radiosonde data. Inset in **a**, average trend profiles of the region 58° N–82° N, 100° W–25° E for full ERA-40 (solid line), subsampled ERA-40 (dashed) and radiosonde data (dotted).

missing because of inconsistent reporting) have their strongest trend near the ground, not above the boundary layer as in the full reanalysis. This is important because boundary layer processes are much more locally driven and simultaneously not well represented in a reanalysis. The same result is found when analysing a subregion with relatively even radiosonde coverage (inset, Fig. 1a), and during the remainder of the year (not shown).

Arctic climate is controlled by processes operating on scales from local to global, including transport effects; forcings such as greenhouse gases, aerosols and clouds; and feedbacks such as the well-known sea-ice–albedo feedback. The temperature profile can be a clue to the underlying processes, but to disentangle the contributions to Arctic temperature trends fully, vertical temperature structures should be addressed in a regionally and seasonally resolved manner. Furthermore, the large interannual variability in the Arctic, coupled with the sensitivity of trends to both end points and season definitions, suggests care should be taken in interpreting trends over short periods.

In conclusion, some features of the temperature trends calculated in ref. 1 reflect possible inhomogeneities or artefacts in the ERA-40 reanalysis rather than true climate signals, as they appear not to be supported by observations. ERA-40 reanalysis is a valuable tool in calculating circulation effects, especially on a subdecadal basis, but inhomogeneities and gaps in the global observing system tend to make trends from reanalyses unreliable, particularly in data-sparse regions.

A. N. Grant[1], **S. Brönnimann**[1] **& L. Haimberger**[2]
[1]Institute for Atmospheric and Climate Science, ETH Zurich, Universitätstrasse 16, CH-8092 Zurich, Switzerland.
e-mail: andrea.grant@env.ethz.ch
[2]Department of Meteorology and Geophysics, University of Vienna, Althanstrasse 14, A-1090 Vienna, Austria.

Received 23 January; accepted 12 May 2008.

1. Graversen, R. G., Mauritsen, T., Tjernström, M., Källén, E. & Svensson, G. Vertical structure of recent Arctic warming. *Nature* **451**, 53–56 (2008).
2. Uppala, S. M. *et al.* The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.* **131**, 2961–3012 (2005).
3. Fu, Q., Johanson, C. M., Warren, S. G. & Seidel, D. J. Contribution of stratospheric cooling to satellite-inferred tropospheric temperature trends. *Nature* **429**, 55–58 (2004).
4. Mears, C. A. & Wentz, F. J. The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science* **309**, 1548–1551 (2005).
5. Bromwich, D. & Wang, S.-H. Evaluation of the NCEP-NCAR and ECMWF 15- and 40-yr reanalyses using rawindsonde data from two independent Arctic field experiments. *Mon. Weath. Rev.* **133**, 3562–3578 (2005).
6. Durre, I., Vose, R. & Wuertz, D. Overview of the integrated global radiosonde archive. *J. Clim.* **19**, 53–68 (2006).
7. Haimberger, L. Homogenization of radiosonde temperature time series using innovation statistics. *J. Clim.* **20**, 1377–1403 (2007).
8. Haimberger, L. *et al.* Towards elimination of the warm bias in historic radiosonde temperature records — some new results from a comprehensive intercomparison of upper air data. *J. Clim.* (in the press).

doi:10.1038/nature07257

Arctic warming aloft is data set dependent

Arising from: R. G. Graversen, T. Mauritsen, M. Tjernström, E. Källén & G. Svensson *Nature* 451, 53–56 (2008)

Arctic sea ice and snow on land have retreated polewards at an alarming pace in the past few decades[1]. Such retreat locally amplifies surface warming through a positive feedback, which causes the Arctic surface to warm faster than the rest of the globe. In contrast, ice and snow retreat causes little warming in the atmosphere above when the stable winter atmosphere inhibits vertical heat exchange. We therefore find surprising the recent report by Graversen *et al.*[2] in which they claim that recent Arctic atmospheric warming extends far deeper into the atmosphere than expected, and can even exceed the surface warming during the polar night. Using a different data set, we show that there is much less warming aloft in winter, consistent with the recent retreat of ice and snow, as well as recent changes in atmospheric heat transport.

Graversen *et al.*[2] compute trends for 1979–2001 from ERA-40 reanalysis, which is a hybrid product using many types of raw observational data assimilated with a consistent global analysis system. The assimilation compensates for some but not all of the variations in the observing system over time that may compromise the veracity of the temperature trend analyses[3]. Figure 1 compares temperature trends in winter from the ERA-40 reanalysis with climate-quality records from satellite observations[4]. Trends in the Arctic winter aloft are strongly data set dependent: the observed trend is 75% less than reanalysis in the middle troposphere and 40% less than in the lower-middle troposphere. In comparison with the observations, the reanalysis exaggerates polar amplification aloft by overestimating the Arctic atmospheric warming and underestimating the Northern

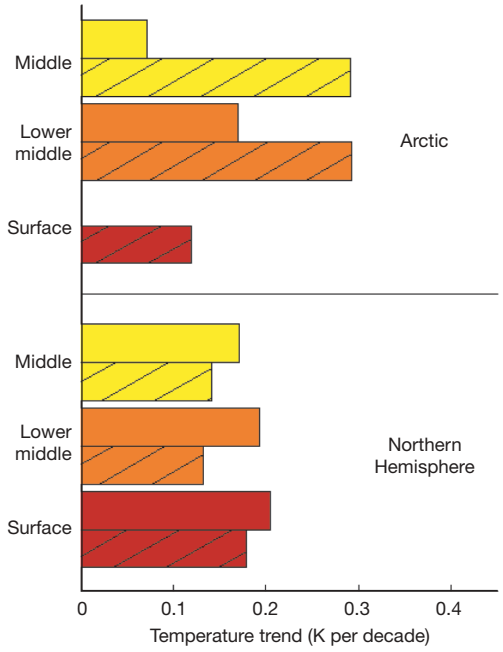


Figure 1 | Temperature trends over the Northern Hemisphere (0°–82.5° N) and the Arctic (65° N–82.5° N) for 1979–2001. Trends are for temperatures at the surface and in the lower-middle and middle troposphere from the ERA-40 reanalysis (hatched) and observations (solid) in the winter (December–February) season. The observed trends are derived from the HadCrut3v (ref. 8) data set for the surface temperature and from a satellite microwave sounding unit[9] (MSU; RSS version 3) for the temperatures in the lower-middle[10] and middle[11,12] troposphere. Observed surface temperatures in the Arctic are not shown, because they are spatially incomplete. For a direct comparison with the MSU observations, synthetic temperatures in the lower-middle and middle troposphere are computed from the ERA-40 reanalysis by applying the MSU weighting functions[3].

Hemisphere atmospheric warming in every season. Specifically, for trends in annual means in the reanalysis for 1979–2001, the Arctic warms 2.7 times more than the Northern Hemisphere in the lower-middle troposphere, in comparison with just 1.5 times more in the observations.

During the polar night, solar absorption at the surface is absent or weak. At the same time, the atmosphere transports a substantial amount of heat northwards from lower latitudes, with heating rates in the Arctic that maximize at about 1,500 m in winter[5]. For these reasons and others, strong radiative cooling at the surface causes frequent lower-tropospheric temperature inversions, which are very stable and damp vertical heat transfer during the polar night. When ice and snow retreat, some of the heat from increased solar absorption is stored at the ocean surface and is released during the cold seasons without warming the atmosphere aloft very much.

It has been concluded that northwards atmospheric heat transport into the Arctic should increase in a warming world[6,7] owing to increased evaporation in the tropics and subsequent condensation in the high latitudes. This increase in latent heat transport is somewhat counterbalanced by a decrease in sensible heat transport, as Arctic amplification decreases the pole-to-equator temperature gradient. Models indicate that warming aloft would not outpace the surface warming after considering increased northwards atmospheric heat transport along with the retreat of ice and snow[7]. Graversen *et al.*[2] find that the change in northwards atmospheric heat transport is not a substantial source of heating aloft in midwinter (January–February) in the Arctic.

The smaller warming trends aloft in the observations in winter are more consistent with the amplification of surface warming from ice and snow retreat and the lack of change in the northwards atmospheric heat transport for 1979–2001. This consistent set of observations calls into question the results of Graversen *et al.*[2] obtained for the polar night.

Cecilia M. Bitz[1] & Qiang Fu[1]

[1]Atmospheric Science Department, 408 Atmospheric Sciences and Geophysics Hall, University of Washington, Seattle, Washington 98195, USA.
e-mail: bitz@atmos.washington.edu

Received 5 February; accepted 9 May 2008.

1. Lemke, P. *et al.* in *Climate Change 2007: The Physical Science Basis* (eds Solomon, S. *et al.*) 337–383 (Contribution of Working Group I to the Fourth Assessment Report of the IPCC, Cambridge Univ. Press, 2007).
2. Graversen, R. G., Mauritsen, T., Tjernström, M., Källén, E. & Svensson, G. Vertical structure of recent Arctic warming. *Nature* **451,** 53–56 (2008).
3. Johanson, C. M. & Fu, Q. Antarctic atmospheric temperature trend patterns from satellite observations. *Geophys. Res. Lett.* **34,** doi:10.1029/2006GL029108 (2007).
4. Karl, T. R., Hassol, S. J., Miller, C. D. & Murray, W. L. (eds) *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences* (Synthesis and Assessment Product 1.1, US Climate Change Science Program, 2006).
5. Overland, J. E. & Turret, P. in *The Polar Oceans and Their Role in Shaping the Global Environment* (eds Johannessen, O. M., Muench, R. & Overland, J. E.) 313–325 (American Geophysical Union, 1994).
6. Alexseev, V. A., Langen, P. L. & Bates, J. R. Polar amplification of surface warming on an aquaplanet in "ghost forcing" experiments without sea ice feedbacks. *Clim. Dyn.* **24,** 655–666 (2005).
7. Cai, M. & Lu, J. Dynamical greenhouse-plus feedback and polar warming amplification. Part II: meridional and vertical asymmetries of the global warming. *Clim. Dyn.* **29,** 375–391 (2007).
8. Brohan, P., Kennedy, J. J., Haris, I., Tett, S. F. B. & Jones, P. D. Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *J. Geophys. Res.* **111,** doi:10.1029/2005JD006548 (2006).
9. Mears, C. A., Schabel, M. C. & Wentz, F. J. A reanalysis of the MSU channel 2 tropospheric temperature record. *J. Clim.* **16,** 3650–3664 (2003).
10. Mears, C. A. & Wentz, F. J. The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science* **309,** 1548–1551 (2005).
11. Fu, Q., Johanson, C. M., Warren, S. G. & Seidel, D. J. Contribution of stratospheric cooling to satellite-inferred tropospheric temperature trends. *Nature* **429,** 55–58 (2004).
12. Johanson, C. M. & Fu, Q. Robustness of tropospheric temperature trends from MSU channels 2 and 4. *J. Clim.* **19,** 4234–4242 (2006).

doi:10.1038/nature07258

Graversen et al. *reply*

Replying to: P. W. Thorne *Nature* 455, doi:10.1038/nature07256; A. N. Grant, S. Brönnimann & L. Haimberger *Nature* 455, doi:10.1038/nature07257; C. M. Bitz & Q. Fu *Nature* 455, doi:10.1038/nature07258 (2008)

These three communications[1–3] question the validity of some of our conclusions[4]. We found Arctic temperature trend amplification well above the boundary layer. In summer, the maximum amplification is found at a height of around 2 km, and no amplification is encountered near the surface. These findings appear in two state-of-the-art reanalyses, ERA-40 (ref. 5) and JRA-25 (ref. 6). Both these data sets show roughly the same overall vertical structure, and we believe our conclusions can be based on either of them. However, they show

considerable differences regarding the magnitudes of the Arctic trends (see our Supplementary Information[4]), but our conclusions are not based on the absolute magnitudes.

A reanalysis synthesizes all available observations and uses a physically based model of the atmosphere to weigh the observations against each other and to extrapolate the observed information in space and time to unobserved parts of the atmosphere. The assimilation procedure takes observational as well as model uncertainties into account. In ERA-40 and JRA-25, the strongest observational constraint on the Arctic temperatures aloft is provided by assimilation of satellite observations, such as microwave sounding unit (MSU) radiances, as *in situ* observations are few in this region. In the assimilation process, careful bias adjustment has been applied to the satellite observations[5].

We examined the agreement of the MSU satellite observations (RSS analysis[7] TLT v3.1 and TMT v3.2) with the vertical structure of ERA-40 and JRA-25. Arctic amplification is encountered in the channel representing the lower troposphere. In summer, both the lower-troposphere and the middle-troposphere channels indicate considerable warming over the Arctic. Because the Arctic surface temperature is constrained to be close to the melting point during this season, this warming must occur aloft, in accordance with the two reanalyses.

The annual lower-troposphere MSU trend reaches 0.46 K per decade at 81.25° N, calculated on the basis of a least-squares fit. We therefore find it surprising that Thorne[1] estimates a high-latitude trend of only 0.2 K per decade. Bitz and Fu[3] report winter trends in the lower troposphere of around 0.2 K per decade both for the Arctic and the Northern Hemisphere, which we also find. However, they report middle-troposphere trends of around 0.09 and 0.18 K per decade for the Arctic and the Northern Hemisphere, respectively. We find, on the other hand, 0.14 and 0.09 K per decade for the Arctic and the Northern Hemisphere, respectively. Hence, the MSU data show winter Arctic amplification in agreement with ERA-40 and JRA-25.

In their last paragraph, Bitz and Fu[3] indicate that ERA-40 exaggerates winter trends aloft. This might be the case; JRA-25 shows considerably smaller trends. However, our point is that, even in JRA-25, winter trends above the boundary layer are comparable to those near the surface and can hardly be linked to surface processes alone. Grant *et al.*[2] compare ERA-40 data with radiosonde observations, which are few in the Arctic. Although these observations cannot confirm the April–October warming aloft found in ERA-40, in general they show good agreement with the ERA-40 data at the points where radiosondes are available.

There is no doubt that more *in situ* observations in the Arctic are needed to enhance the quality of future reanalyses. Given the absence of such observations in historical archives, we feel that a reanalysis is likely to provide a better representation of the true state of the Arctic atmosphere than any single inhomogeneous set of a specific observation type. Satellite observations must be bias corrected and radio soundings exist almost only in the southern part of the Arctic. In a reanalysis, both of these shortcomings are consistently handled in the framework of a dynamical, global model of the atmosphere. We have given an estimate of the uncertainty associated with reanalysis data by displaying results from two different, second-generation reanalyses. Within the limits of this uncertainty we believe that our conclusions remain valid.

R. G. Graversen[1], T. Mauritsen[1], M. Tjernström[1], E. Källén[1] & G. Svensson[1]

[1]Department of Meteorology, Stockholm University, S-106 91 Stockholm, Sweden.

e-mail: rune@misu.su.se

1. Thorne, P. W. Arctic tropospheric warming amplification? *Nature* **455,** doi:10.1038/nature07256 (2008).
2. Grant, A. N., Brönnimann, S. & Haimberger, L. Recent Arctic warming vertical structure contested. *Nature* **455,** doi:10.1038/nature07257 (2008).
3. Bitz, C. M. & Fu, Q. Arctic warming aloft is data set dependent. *Nature* **455,** doi:10.1038/nature07258 (2008).
4. Graversen, R. G., Mauritsen, T., Tjernström, M., Källén, E. & Svensson, G. Vertical structure of recent Arctic warming. *Nature* **451,** 53–56 (2008).
5. Uppala, S. M. *et al.* The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.* **131,** 2961–3012 (2005).
6. Onogi, K. *et al.* The JRA-25 reanalysis. *J. Meteorol. Soc. Jpn* **85,** 369–432 (2007).
7. Mears, C. A. & Wentz, F. J. The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science* **309,** 1548–1551 (2005).

doi:10.1038/nature07259

ARTICLES

Broadband observations of the naked-eye γ-ray burst GRB 080319B

J. L. Racusin[1], S. V. Karpov[2], M. Sokolowski[3], J. Granot[4], X. F. Wu[1,5], V. Pal'shin[6], S. Covino[7], A. J. van der Horst[8], S. R. Oates[9], P. Schady[9], R. J. Smith[10], J. Cummings[11], R. L. C. Starling[12], L. W. Piotrowski[13], B. Zhang[14], P. A. Evans[12], S. T. Holland[15,16,17], K. Malek[18], M. T. Page[9], L. Vetere[1], R. Margutti[19], C. Guidorzi[7,10], A. P. Kamble[20], P. A. Curran[20], A. Beardmore[12], C. Kouveliotou[21], L. Mankiewicz[18], A. Melandri[10], P. T. O'Brien[12], K. L. Page[12], T. Piran[22], N. R. Tanvir[12], G. Wrochna[3], R. L. Aptekar[6], S. Barthelmy[11], C. Bartolini[23], G. M. Beskin[2], S. Bondar[24], M. Bremer[25], S. Campana[7], A. Castro-Tirado[26], A. Cucchiara[1], M. Cwiok[13], P. D'Avanzo[7], V. D'Elia[27], M. Della Valle[28,29], A. de Ugarte Postigo[30], W. Dominik[13], A. Falcone[1], F. Fiore[27], D. B. Fox[1], D. D. Frederiks[6], A. S. Fruchter[31], D. Fugazza[7], M. A. Garrett[32,33,34], N. Gehrels[11], S. Golenetskii[6], A. Gomboc[35], J. Gorosabel[26], G. Greco[23], A. Guarnieri[23], S. Immler[15,17], M. Jelinek[26], G. Kasprowicz[36], V. La Parola[37], A. J. Levan[38], V. Mangano[37], E. P. Mazets[6], E. Molinari[7], A. Moretti[7], K. Nawrocki[3], P. P. Oleynik[6], J. P. Osborne[12], C. Pagani[1], S. B. Pandey[39], Z. Paragi[40], M. Perri[41], A. Piccioni[23], E. Ramirez-Ruiz[42], P. W. A. Roming[1], I. A. Steele[10], R. G. Strom[20,32], V. Testa[27], G. Tosti[43], M. V. Ulanov[6], K. Wiersema[12], R. A. M. J. Wijers[20], J. M. Winters[25], A. F. Zarnecki[13], F. Zerbi[7], P. Mészáros[1,44], G. Chincarini[7,19] & D. N. Burrows[1]

Long-duration γ-ray bursts (GRBs) release copious amounts of energy across the entire electromagnetic spectrum, and so provide a window into the process of black hole formation from the collapse of massive stars. Previous early optical observations of even the most exceptional GRBs (990123 and 030329) lacked both the temporal resolution to probe the optical flash in detail and the accuracy needed to trace the transition from the prompt emission within the outflow to external shocks caused by interaction with the progenitor environment. Here we report observations of the extraordinarily bright prompt optical and γ-ray emission of GRB 080319B that provide diagnostics within seconds of its formation, followed by broadband observations of the afterglow decay that continued for weeks. We show that the prompt emission stems from a single physical region, implying an extremely relativistic outflow that propagates within the narrow inner core of a two-component jet.

The GRB 080319B, discovered by NASA's Swift GRB Explorer mission[1] on 19 March 2008, set new records among these most luminous transient events in the Universe. GRBs are widely thought to occur through the ejection of a highly relativistic, collimated outflow (jet), produced by a newly formed black hole. Under the standard fireball model[2–6], collimated relativistic shells propagate away from the central engine, crash into each other (internal shocks) and decelerate as they plough into the surrounding medium (external/forward

[1]Department of Astronomy and Astrophysics, 525 Davey Laboratory, Pennsylvania State University, University Park, Pennsylvania 16802, USA. [2]Special Astrophysical Observatory, Nizhnij Arkhyz, Zelenchukskij region, Karachai-Cirkassian Republic, Russia 369167. [3]Soltan Institute for Nuclear Studies, 05-400 Otwock-Swierk, Poland. [4]Centre for Astrophysics Research, University of Hertfordshire, College Lane, Hatfield AL10 9AB, UK. [5]Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing 210008, China. [6]Ioffe Physico-Technical Institute, Laboratory for Experimental Astrophysics, Saint Petersburg 194021, Russian Federation. [7]INAF-Osservatorio Astronomico di Brera, via E. Bianchi 46, I-23807 Merate (LC), Italy. [8]NASA Postdoctoral Program Fellow, NSSTC, 320 Sparkman Drive, Huntsville, Alabama 35805, USA. [9]The UCL Mullard Space Science Laboratory, Holmbury St Mary, Surrey RH5 6NT, UK. [10]Astrophysics Research Institute, Liverpool John Moores University, Twelve Quays House, Birkenhead CH41 1LD, UK. [11]Astrophysics Science Division, Code 661, NASA's Goddard Space Flight Centre, 8800 Greenbelt Road, Greenbelt, Maryland 20771, USA. [12]Department of Physics and Astronomy, University of Leicester, Leicester LE1 7RH, UK. [13]Institute of Experimental Physics, University of Warsaw, Hoża 69, 00-681 Warsaw, Poland. [14]Department of Physics and Astronomy, University of Nevada Las Vegas, Las Vegas, Nevada 89154, USA. [15]Astrophysics Science Division, Code 660.1, NASA Goddard Space Flight Centre, 8800 Greenbelt Road, Greenbelt, Maryland 20771, USA. [16]Universities Space Research Association, 10211 Wincopin Circle, Suite 500, Columbia, Maryland 21044, USA. [17]Centre for Research and Exploration in Space Science and Technology, Code 668.8, NASA Goddard Space Flight Centre, 8800 Greenbelt Road, Greenbelt, Maryland 20771, USA. [18]Center for Theoretical Physics PAS, Al. Lotników 32/46, 02-668 Warsaw, Poland. [19]Università degli Studi di Milano Bicocca, Physics Department, Piazza della Scienza 3, I-20126 Milano, Italy. [20]Astronomical Institute 'Anton Pannekoek', University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands. [21]NASA/Marshall Space Flight Center, VP62, NSSTC, 320 Sparkman Drive, Huntsville, Alabama 35805, USA. [22]Racah Institute for Physics, The Hebrew University, Jerusalem, 91904, Israel. [23]Universitá di Bologna, Via Ranzani 1, 40126 Bologna, Italy. [24]Institute for Precise Instrumentation, Nizhnij Arkhyz 369167, Russian Federation. [25]Institute de Radioastronomie Millimétrique (IRAM), 300 rue de la Piscine, 38406 Saint Martin d'Hères, France. [26]Instituto de Astrofísica de Andalucía (IAA-CSIC), PO Box 03004, 18080 Granada, Spain. [27]INAF – Osservatorio Astronomico di Roma, via Frascati 33, 00040 Monteporzio Catone, Italy. [28]European Southern Observatory, Karl-Schwarzschild-Strasse 2, D-85748 Garching bei München, Germany. [29]INAF – Osservatorio Astronomico di Capodimonte, Salita Moiariello, 16, 80131 Napoli, Italy. [30]European Southern Observatory, Casilla 19001, Santiago 19, Chile. [31]Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, Maryland 21218, USA. [32]Netherlands Institute for Radio Astronomy (ASTRON), Postbus 2, 7990 AA Dwingeloo, The Netherlands. [33]Leiden Observatory, University of Leiden, PB 9513, Leiden 2300 RA, The Netherlands. [34]Centre for Astrophysics and Supercomputing, Swinburne University of Technology, Hawthorn, Victoria 3122, Australia. [35]Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 19, SI-1000 Ljubljana, Slovenia. [36]Institute of Electronic Systems, Warsaw University of Technology, Nowowiejska 15/19, 00-665 Warsaw, Poland. [37]INAF – IASF PA, Via Ugo La Malfa 153, 90146 Palermo, Italy. [38]Department of Physics, University of Warwick, Coventry CV4 7AL, UK. [39]Aryabhatta Research Institute of Observational-Sciences (ARIES), Manora Peak, NainiTal, Uttarakhand 263129, India. [40]Joint Institute for VLBI in Europe (JIVE), Postbus 2, 7990 AA Dwingeloo, The Netherlands. [41]ASI Science Data Center, c/o ESRIN, via G. Galilei, 00044 Frascati, Italy. [42]Department of Astronomy and Astrophysics, University of California, Santa Cruz, California 95064, USA. [43]University of Perugia, Piazza dell'Università 1, 06100 Perugia, Italy. [44]Physics Department, 104 Davey Laboratory, Pennsylvania State University, University Park, Pennsylvania 16801, USA.

shocks). Reverse shocks propagate back into the jet, generating optical emission. With a uniquely bright peak visual magnitude of 5.3 (Fig. 1) at a redshift of $z = 0.937$ (ref. 7), GRB 080319B was the brightest optical burst ever observed. An observer in a dark location could have seen the prompt optical emission with the naked eye. The astronomical community has been waiting for such an event for the past nine years, ever since GRB 990123 (the previous record holder for the highest peak optical brightness) peaked at a visual magnitude of ~9, leading to significant insight into the GRB optical emission mechanisms[8].

The location of GRB 080319B was fortuitously only 10° away from GRB 080319A, detected by Swift less than 30 min earlier, allowing several wide-field telescopes to detect the optical counterpart of GRB 080319B instantly. The rapid localization by Swift enabled prompt multi-wavelength follow-up observations by robotic ground-based telescopes, resulting in arguably the best broadband GRB observations obtained so far. These observations continued for weeks afterwards as we followed the fading afterglow, providing strong constraints on the physics of the explosion and its aftermath.

At its peak, GRB 080319B displayed the brightest optical and X-ray fluxes ever measured for a GRB, and one of the highest γ-ray fluences recorded. Our broadband data cover 11.5 orders of magnitude in wavelength, from radio to γ-rays, and begin (in the optical and γ-ray bands) before the explosion. We identify three different components responsible for the optical emission. The earliest data (at $t \equiv T - T_0 < 50$ s) provide evidence that the bright optical and γ-ray emissions stem from the same physical region within the outflow. The second optical component ($50\,\mathrm{s} < t < 800$ s) shows the distinct characteristics of a reverse shock, and the final component (at $t > 800$ s) represents the afterglow produced as the external forward shock propagates into the surrounding medium. Previous measurements of GRBs have revealed one or two of these components at a time[9–11], but never all three in the same burst with such clarity. GRB 080319B is therefore a testbed for broad theoretical modelling of GRBs and their environments.

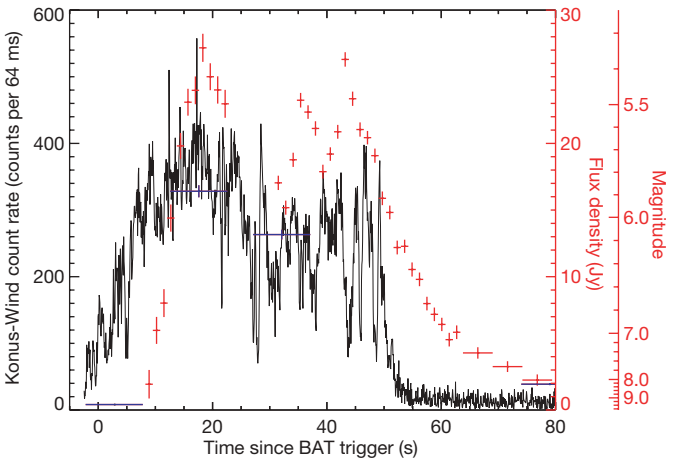


Figure 1 | Light curve of prompt emission. The Konus-Wind background-subtracted γ-ray light curve (black; 18–1,160 keV), shown relative to the trigger time T_0 of the Swift-BAT. The burst had a peak γ-ray flux of $F_p = (2.26 \pm 0.21) \times 10^{-5}$ erg cm^{-2} s^{-1}, a fluence F_γ (20 keV to 7 MeV) of $(6.23 \pm 0.13) \times 10^{-4}$ erg cm^{-2}, a peak isotropic equivalent luminosity $L_{\mathrm{p,iso}}$ of $(1.01 \pm 0.09) \times 10^{53}$ erg s^{-1} (at the luminosity distance d_{L} of 1.9×10^{28} cm, assuming cosmological parameters $H_0 = 71$ km s^{-1} Mpc^{-1}, $\Omega_{\mathrm{M}} = 0.27$ and $\Omega_\Lambda = 0.73$), and an isotropic equivalent γ-ray energy release $E_{\gamma,\mathrm{iso}}$ of 1.3×10^{54} erg (20 keV to 7 MeV). These are among the highest measured so far. Optical data from 'Pi of the Sky' (blue) and TORTORA (red) are superimposed for comparison. The optical emission begins within seconds of the onset of the burst. The TORTORA data have a gap during the slew of the REM telescope to this field, but show three subpeaks in the optical brightness, reaching a peak brightness of 5.3 mag (white). The γ-ray light curve has multiple short peaks; these are not positively correlated with the optical peaks in detail (compare with ref. 23). If the synchrotron self-absorption frequency is slightly above the optical emission, this may account for the broad optical peaks and the lack of detailed correlation. However, the optical flash begins and ends at about the same times, providing strong evidence that both originate at the same site. See Supplementary Information for a more detailed description of correlation tests. All plotted error bars are 1σ, and quoted parameter errors are 90% confidence.

Discovery and broadband observations

Swift's Burst Alert Telescope (BAT[12]; 15–350 keV) triggered[13] on GRB 080319B at $T_0 = 06{:}12{:}49$ UT on 19 March 2008. The burst was detected simultaneously with the Konus γ-ray detector (20 keV to 15 MeV) on board the Wind satellite[14,15]. Both the BAT and Konus-Wind (KW) light curves (Supplementary Figs 1 and 3) show a complex, strongly energy-dependent structure, with many clearly separated pulses above 70 keV and a generally smoother behaviour at lower energies, lasting ~57 s.

The wide-field robotic optical telescope 'Pi of the Sky'[16] and the wide-field robotic instrument Telescopio Ottimizzato per la Ricerca dei Transienti Ottici Rapidi (TORTORA[17]) both serendipitously had the GRB within their fields of view at the time of the explosion (as they were both already observing GRB 080319A (ref. 18)). 'Pi of the Sky' observed the onset of the bright optical transient, which began at 2.75 ± 5 s after the BAT trigger, rose rapidly, peaked at $\sim T_0 + 18$ s and then faded below the threshold to magnitude ~12 after 5 min. TORTORA measured the brightest portion of the optical flash with high time resolution, catching three separate peaks (Fig. 1) and enabling us to do detailed comparisons between the prompt optical and γ-ray emissions.

The Swift spacecraft and the Rapid Eye Mount (REM[19]) telescope both initiated automatic slews to the burst, resulting in optical observations with REM and the Swift Ultraviolet–Optical Telescope (UVOT[20]), and X-ray observations with the Swift X-ray Telescope (XRT[21]). Over the next several hours we obtained ultraviolet, optical and near-infrared (NIR) photometric observations of the GRB afterglow with the Swift-UVOT, REM, the Liverpool Telescope, the Faulkes Telescope North, Gemini-North, and the Very Large Telescope. Subsequent optical spectroscopy by Gemini-N and the Hobby–Eberly Telescope confirmed the redshift of 0.937 (Supplementary Figs 4 and 5). A millimetre-wavelength counterpart was detected with the IRAM Plateau de Bure Interferometer at $\sim T_0 + 16$ h. Multiple epochs of radio observations with the Westerbork Synthesis Radio Telescope revealed a radio counterpart ~2–3 days after the burst. X-ray and optical observations continued for more than four weeks after the burst. The composite broadband light curves of GRB 080319B, which include all data discussed throughout this paper, and cover eight orders of magnitude in flux and more than six orders of magnitude in time, are shown in Fig. 2 and summarized in Table 1. All of these data are given in Supplementary Information.

Ultra-relativistic prompt emission

The contemporaneous bright 'optical flash' and the γ-ray burst (Fig. 1) provide important constraints on the nature of the prompt GRB emission mechanism. Although there is a general consensus that the prompt γ-rays must arise from internal dissipation within the outflow, probably as a result of internal shocks, the optical flash may arise either from the same emitting region as the γ-rays or from the reverse shock that decelerates the outflow as it sweeps up the external medium. The reverse shock becomes important when the inertia of the swept-up external matter starts to slow down the ejecta appreciably, at a larger radius than the dissipation by internal shocks.

The temporal coincidence of the onset and overall shape of the prompt optical and γ-ray emissions suggest that both originate from the same physical region (see also refs 22, 23), although their respective peaks during this phase do not positively correlate in detail (see Supplementary Figs 8–10 and the related discussion in Supplementary Information). Nevertheless, the initial steep rise (at

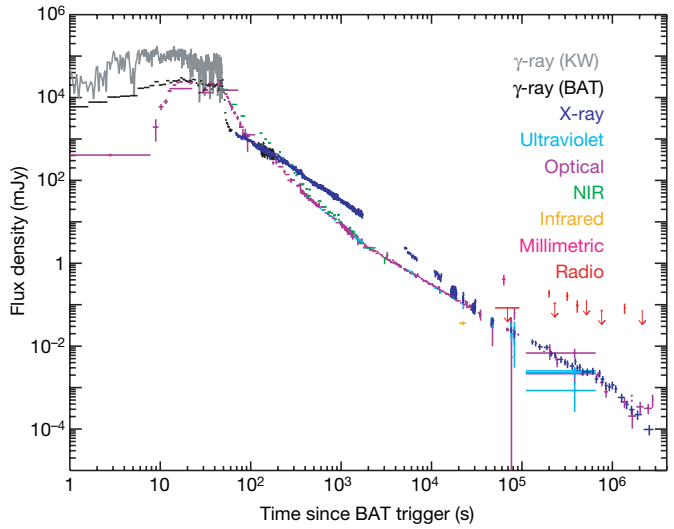


Figure 2 | Composite light curve. Broadband light curve of GRB 080319B, including radio, millimetric, infrared, NIR, optical, ultraviolet, X-ray and γ-ray flux densities. The ultraviolet, optical and NIR data are normalized to the UVOT v-band in the interval between $T_0 + 1{,}500$ s and $T_0 + 10$ ks. The Swift-BAT data are extrapolated down into the XRT bandpass (0.3–10 keV) for direct comparison with the XRT data. The combined X-ray and BAT data are scaled up by a factor of 45, and the Konus-Wind (KW) data are scaled up by a factor of 10^4 for comparison with the optical flux densities. This figure includes our own data, plus one VLA radio data point[49], and optical data from the Katzman Automatic Imaging Telescope, Nickel and Gemini-S[22]. The deviations in the NIR points from $T_0 + 100$ s to $T_0 + 600$ s are due to strong colour evolution in the spectral energy distributions at this time; these points were not included in our overall light-curve fits (Supplementary Fig. 6). After the optical flash, the optical light curve is best described by the superposition of three different power-law components (Supplementary Fig. 6) with decay indices of $\alpha_{opt,1} = 6.5 \pm 0.9$ (the tail of the optical flash), $\alpha_{opt,2} = 2.49 \pm 0.09$ and $\alpha_{opt,3} = 1.25 \pm 0.02$. The X-ray light curve clearly differs from the optical light curve during the first ~12 h. After a short flat smooth transition from the tail of the γ-ray prompt emission, the X-ray light curve (Supplementary Fig. 7) after ~80 s can be fitted by a triple broken power-law function with decay indices of 1.44 ± 0.07, 1.85 ± 0.10, $1.17^{+0.14}_{-0.23}$ and $2.61^{+2.04}_{-0.91}$, and with break times of $2{,}242 \pm 940$ s, $4.1^{+2.8}_{-1.7} \times 10^4$ s and $(1.0 \pm 0.5) \times 10^6$ s (χ^2/d.f. = 880/697 = 1.26), or by the superposition of two broken power-laws with decay indices of 1.45 ± 0.05, $2.05^{+0.44}_{-0.27}$, $0.95^{+0.19}_{-0.69}$ and $2.70^{+2.06}_{-1.12}$, and break times of $2{,}800^{+900}_{-1{,}400}$ s and $9.5^{+6.2}_{-4.1} \times 10^5$ s (χ^2/d.f. = 902/701 = 1.29). All plotted error bars are 1σ, and quoted parameter errors are 90% confidence.

$t < 18$ s), the rapid decline (at $t > 43$ s) and the constant optical pulse widths indicate[24–26] that the optical flash did not arise from a reverse shock (compare with GRB 990123; refs 27, 28).

The flux density of the optical flash is $\sim 10^4$ times larger than the extrapolation of the γ-ray spectrum into the optical band (Fig. 3). The popular interpretation of the soft γ-rays as synchrotron emission cannot account for such a bright optical component from the same physical region, suggesting that different radiation mechanisms must dominate in each spectral regime. The most natural (but by no means the only viable) candidates are synchrotron for the optical component and synchrotron self-Compton (SSC) for the γ-rays[29,30]. The Compton Y parameter, defined as the ratio of the inverse Compton to synchrotron energy losses, is $Y \sim \nu F_\nu(E_p)/\nu F_\nu(E_{p,syn}) \gtrsim 10$, where $E_{p,syn}$ is the peak photon energy of the synchrotron νF_ν spectrum, to account for the fact that the prompt γ-ray energy is higher than the prompt optical/ultraviolet synchrotron energy. This would imply a third spectral component arising from second-order inverse Compton scattering that peaks at energies around $E_{p,2} \approx E_p^2/E_{p,syn} \approx 23(E_{p,syn}/20\,\mathrm{eV})^{-1}$ GeV. Note that the Klein–Nishina suppression becomes important only at $E > 94(E_{p,syn}/20\,\mathrm{eV})^{-1/2}\Gamma_3$ GeV, where $\Gamma = 10^3\Gamma_3$ is the outflow bulk Lorentz factor. This third spectral component carries more energy than the observed γ-rays, by a factor $Y \gtrsim 10$, changing the energy budget of this burst and implying that GRB 080319B was even more powerful than inferred from the observed emission. Most of the energy in this burst was emitted in this undetected GeV component, which would have been detected by the Astro-rivelatore Gamma a Immagini Leggero (AGILE) satellite had it not been occulted by the Earth, and would have been easily detectable by the recently launched Gamma-ray Large Area Space Telescope (GLAST)[31] satellite.

Such bright prompt optical flashes are rare. The exceptional brightness of the optical flash in GRB 080319B implies that the self-absorption frequency ν_a cannot be far above the optical band near the peak time. The optical brightness temperature implies that $300 \leq \Gamma(t_v/3\,\mathrm{s})^{2/3} \leq 1400$, and therefore $\Gamma \sim 10^3$, where $t_v \equiv R\Gamma^{-2}c^{-1}$ is the rough variability timescale in the internal shocks model. Because of the extremely high bulk Lorentz factor Γ, the internal shocks occur at an unusually large radius given by $0.8 \leq R_{16}(t_v/3\,\mathrm{s})^{1/3} \leq 20$, where $R_{16} = R/10^{16}$ cm, resulting in a relatively low ν_a, which in turn allows the optical photons to escape.

Interpretation of the chromatic afterglow

Our broadband data set enabled us to measure the temporal and spectral evolution of GRB 080319B throughout the afterglow. After the prompt phase, the early (minutes to hours) X-ray and optical behaviour are inconsistent with the predictions of the standard afterglow theory, suggesting that they must stem from different emission regions. In particular we find that the optical, X-ray and γ-ray emissions from this burst are explained reasonably well by a two-component jet model[32–38] (Fig. 4 and Table 2), consisting of an ultra-relativistic narrow jet, surrounded by a broader jet with a lower Lorentz factor. The empirical triple broken power-law function of the X-ray light curve is then interpreted as the superposition of two broken power-law components representing these two jets (Table 2 and Supplementary Fig. 7). This structure, in which the Lorentz factor and energy per solid angle are highest near the axis and decrease outwards, either smoothly or in quasi-steps, qualitatively resembles the results of numerical simulations of jet formation in collapsars[39]. Further details of the model are given in Supplementary Information; here we summarize the model results and apply them to the observational data.

The optical light curve at $50\,\mathrm{s} < t < 800$ s is dominated by the second optical power-law component, which we interpret as emission from the reverse shock associated with the interaction of the wide jet with the external medium. This segment is consistent with expectations for the high-latitude emission[40] from a reverse shock ($\alpha = 2 + \beta$) if the cooling frequency ν_c is below the optical band and the injection frequency $\nu_m > 10^{16}$ Hz. Emission from the reverse shock peaks at $t \approx 50$ s in the optical with a peak flux density of ~2–3 Jy, but it is initially overwhelmed by the much brighter prompt emission and does not become visible until the latter dies away. The high peak luminosity of the optical reverse shock component soon after the end of the γ-ray emission indicates that the reverse shock was at least mildly relativistic. The GRB outflow could not have been highly magnetized ($\sigma \gg 1$) when it crossed the reverse shock, or the reverse shock would have been suppressed[41], implying that $\sigma \lesssim 1$, where σ is the ratio of electromagnetic to kinetic energy flux. However, if the outflow was too weakly magnetized ($\sigma \ll 1$), the optical emission would also have been suppressed. Therefore an intermediate magnetization ($0.1 \leq \sigma \leq 1$) is needed to obtain the observed bright emission from the reverse shock[42,43].

In contrast, the X-ray light curve in the interval $50\,\mathrm{s} < t < 40$ ks is dominated by the forward shock of the narrow jet component interacting with a surrounding medium produced by the wind[44] of the progenitor star in the slow cooling case ($\nu_m < \nu_X < \nu_c$, where ν_X indicates the X-ray band). The first break in the X-ray light curve is attributed to a jet break[45] in this narrow jet (Table 2), leading to a jet half-opening angle of ~0.2°. Because this break is not seen in the

Table 1 | Observations of GRB 080319B

Facility	Epoch*	Band	Peak flux†
Swift-BAT	−120 to 182	15–350 keV	2.3×10^{-6} erg cm^{-2} s^{-1}
Konus-Wind	−2 to 230	20–1,160 keV‡	2.3×10^{-5} erg cm^{-2} s^{-1}
Swift-XRT	67 to 2.5×10^{6}	0.3–10 keV	–
'Pi of the Sky'	−1,380 to 468	White	5.9 mag
TORTORA	−20 to 97	V	5.3 mag
Swift-UVOT	68–10^{6}	White, u, v, b, uvw1, uvw2, uvm2	–
REM	51–2,070	R, I, J, H, Ks	–
Liverpool Telescope	$(1.8–2.5)\times10^{3}$	SDSS r,i	–
Faulkes Telescope North	$(2.5–20.5)\times10^{4}$	Bessell R,I SDSS r,i	–
Very Large Telescope	435–934	J, Ks	–
Gemini N Photometry	3.0×10^{5}, 4.5×10^{5}	r, i	–
HST	1.6×10^{6}	F606W, F814W	–
Gemini N Spectroscopy	$(1.2–1.24)\times10^{4}$	4,100-6,800 Å	–
Hobby–Eberly Telescope	$(2.0–2.1)\times10^{4}$	4,100-10,500 Å	–
Westerbork Synthesis Radio Telescope	$(5.1–220)\times10^{4}$	4.8 GHz	–
IRAM-Plateau de Bure	$(6.0–6.6)\times10^{4}$	97.98 GHz	–
VLA§	$(1.98–2.02)\times10^{5}$	4.86 GHz	189 μJy
Pairitel§	$(1.27–1.77)\times10^{4}$	J, H, Ks	–
KAIT§	$(0.1–1.7)\times10^{4}$	Clear, B, V, I	–
Nickel§	$(0.7–2.4)\times10^{4}$	B, V, R, I	–
Gemini S§	$(0.9–1.7)\times10^{5}$	g, r, i, z	–
Spitzer§	$(2.20–2.24)\times10^{4}$	15.8 μm	–

Details of our observations and data analysis are given in Supplementary Methods.
* Time since BAT trigger in seconds.
† Peak fluxes listed only if a peak was actually observed.
‡ Konus-Wind light curve measured in the 20–1,160 keV range; peak flux measured in the range 20 keV to 7 MeV.
§ Observations obtained from external sources as identified in Supplementary Methods.

optical light curve, the optical flux from the forward shock of the narrow jet must be much less than that of the wide jet, implying that $\nu_{opt} < \nu_m < \nu_X < \nu_c$ (where ν_{opt} indicates the optical band).

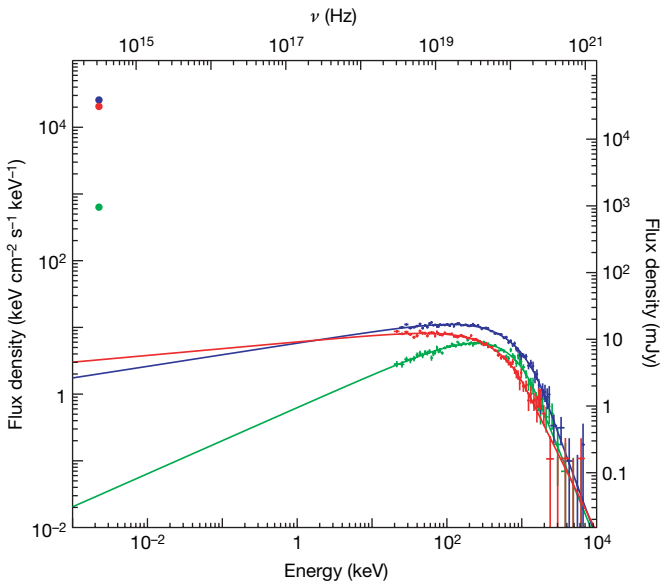


Figure 3 | Spectral energy distribution of the prompt emission. Konus-Wind spectra and 'Pi of the Sky' flux density in three 10-s time intervals centred at $T_0 + 3$ s (green), $T_0 + 17$ s (blue) and $T_0 + 32$ s (red). (Detailed time intervals and γ-ray spectral parameters are given in Supplementary Table 1.) The high-energy data points are from Konus-Wind, and the solid line shows the best-fit Band function[50] for each time interval. The time-resolved Konus-Wind spectra show that the Band-function parameters vary rapidly during the prompt emission, with the low-energy slope changing from −0.5 to −0.9 and E_p changing from ~740 keV to ~540 keV in the first 30 s (see Supplementary Table 1 and Supplementary Fig. 3). Time-resolved single power-law spectral fits of the BAT data show the photon index shifting rapidly from ~1.0 to ~2.1 at $T_0 + 53$ s (near the end of the prompt phase; Supplementary Fig. 2). The low-energy points are the 'Pi of the Sky' flux density measured during about the same time interval. The optical flux density exceeds the extrapolation of the γ-ray model by four orders of magnitude. All plotted error bars are 1σ.

The optical emission after $T_0 + 800$ s is dominated by a single power-law function, which is consistent with the expectation for forward shock emission from the wide jet with $\nu_m < \nu_{opt} < \nu_c$. The late X-ray afterglow after 40 ks is also dominated by the forward shock of the wide jet with an overall spectrum of $\nu_m < \nu_{opt} < \nu_c < \nu_X$. At about 11 days after the burst, the X-ray light curve breaks to a steeper slope (confirmed by a late observation with the Chandra X-ray Observatory; E. Rol, personal communication). If this break is interpreted as the jet break of the wide jet (Table 2), it corresponds to an initial jet half-opening angle of ~4°. The forward shock of the wide jet also accounts for the observed radio emission, which is strongly modulated by the effects of Galactic scintillation (see Supplementary Methods for more detailed discussion)[46,47] when the source is small.

Because the observed γ-ray emission of GRB 080319B shows very similar properties to those of most GRBs, it may be representative of the main underlying physical mechanism. If so, similar lower-energy spectral components would be expected in most GRBs. The paucity of bright optical flashes may be attributed to less relativistic outflows in most GRBs, leading to smaller emitting radii R, higher optical depths, and significantly higher values of ν_a, ultimately suppressing the optical emission. In this model, the spectacular optical brightness of GRB 080318B is due mainly to its unusually large Γ. Previous examples of GRBs with bright optical counterparts[9–11] (990123, 041219a and 050820a) that also had large initial Γ values either lacked the high γ-ray luminosities or resided in a constant density and not a wind environment as with 080319B, suppressing the bright optical flash.

The afterglow may also be interpreted by alternative models such as a blast wave propagating into a complex medium (see Supplementary Figs 14 and 15 and the related discussions in Supplementary Information), or evolving microphysical parameters, but we consider the two-component jet model to be the most plausible interpretation. An interesting consequence of these theoretical considerations is that GRB 080319B, which has the best broadband data set recorded so far, is not consistent with the expectations of any of the simple GRB models previously studied. The case for multiple spectral emission components and the two-component jet presented here suggests that similar models may be able to explain at least some of the chromatic breaks seen in optical and X-ray afterglows over the

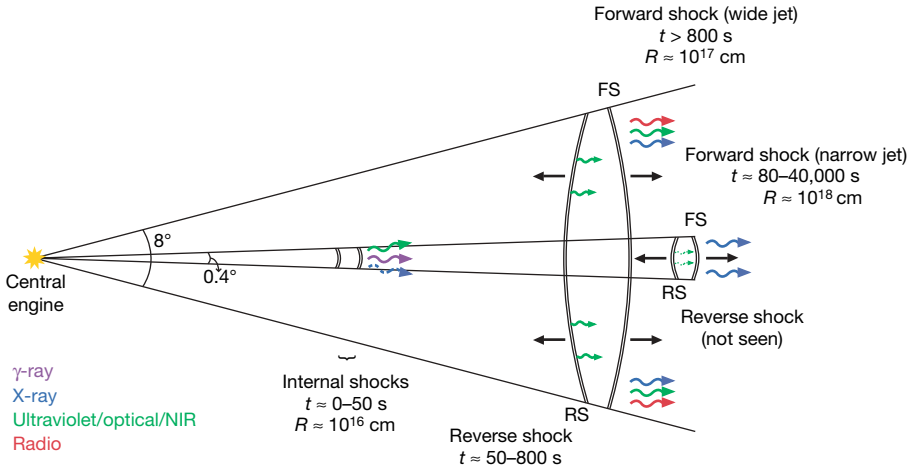


Figure 4 | Schematic diagram of the two-component jet model. This summary diagram shows spectral and temporal elements of our model. The prompt γ-ray emission is due to the internal shocks in the narrow jet, and the afterglow is a result of the forward and reverse shocks from both the narrow and wide jets. The reverse shock from the narrow jet is too faint to detect in comparison with the bright wide-jet reverse shock and the prompt emission. If X-ray observations had begun earlier, we would have detected X-ray emission during the prompt burst. These expected (but unobserved) emission sources are indicated by the dashed photon lines. (Diagram courtesy of J. D. Myers.)

Table 2 | Summary of two-component jet parameters

Parameter	α_{opt}	β_{opt}	α_x	β_x	p	ν_m	ν_c	t_j (s)	θ_j (deg)*	E_γ (erg)
WJRS	2.49 ± 0.09	0.49 ± 0.14	–	–	–	$>\nu_{opt}$; $<\nu_X$	$<\nu_{opt}$	–	–	–
NJFS	–	–	1.45 ± 0.05; $2.05^{+0.44}_{-0.27}$†	0.76 ± 0.10	2.4	$>\nu_{opt}$; $<\nu_X$	$>\nu_X$; $<\nu_\gamma$	$2{,}800^{+900}_{-1{,}400}$	0.2	2.1×10^{50}
WJFS	1.25 ± 0.02	0.50 ± 0.07	0.95 ± 0.20; $2.70^{+2.06}_{-1.12}$†	0.98 ± 0.10	2.0	$<\nu_{opt}$	$<\nu_X$; $>\nu_{opt}$	$9.5^{+6.2}_{-4.1} \times 10^5$	4.0	1.9×10^{50}

The radiation mechanism is assumed to be synchrotron emission, as postulated in the standard fireball model[2–6], and the spectral and light curve segments are fitted by power laws, described by $F_\nu = t^{-\alpha}\nu^{-\beta}$, with decay index α and spectral energy index β. The details of the temporal fits are given in Fig. 2 and Supplementary Figs 6 and 7. We created broadband spectral energy distributions at 11 epochs ranging from $T_0 + 150$ s to $T_0 + 500$ ks. The spectral energy distributions can be fitted by power laws, broken power laws, or double broken power laws. Deviations from broken power laws at the ultraviolet and soft X-ray frequencies allow us to measure the host galaxy extinction and X-ray absorption column density (see discussion in Supplementary Information). Our resulting best-fit spectral models are described in detail in Supplementary Information and shown in Supplementary Figs 11–13. WJRS, wide-jet reverse shock; NJFS, narrow-jet forward shock; WJFS, wide-jet forward shock.

* Half-opening angle.

† Post-jet break temporal slope for each jet.

past few years that have been difficult to reconcile with the standard models[3,4].

The probability of being located within the tiny solid angle of the narrow jet is small ($\sim 10^{-3}$). If every GRB has such a narrow jet, we should expect to detect the narrow-jet emission from a GRB every ~3–10 years. Had we observed GRB 080319B even slightly off-axis, the behaviour might have appeared similar to many other GRB afterglows. Despite the incredibly high flux and fluence of GRB 080319B, the total jet-corrected observed energy budget ($\sim 4 \times 10^{50}$ erg) is moderate and is consistent with the overall distribution for all GRBs[48]. In addition, if the SSC interpretation of the prompt emission is indeed generic, it implies that a reasonably bright second-order SSC component peaking at ~10–100 GeV may be a common feature in GRBs and may significantly increase the total energy budget of a GRB. This GeV emission would be seen with a delay of a few seconds compared with the optical emission. GLAST will soon test these predictions.

Received 11 May; accepted 11 July 2008.

1. Gehrels, N. *et al.* The Swift gamma-ray burst mission. *Astrophys. J.* **611,** L1005–L1020 (2004).
2. Rees, M. J. & Mészáros, P. Relativistic fireballs—energy conversion and time-scales. *Mon. Not. R. Astron. Soc.* **258,** 41–43 (1992).
3. Mészáros, P. & Rees, M. J. Optical and long-wavelength afterglow from gamma-ray bursts. *Astrophys. J.* **476,** 232–237 (1997).
4. Wijers, R. A. M. J., Rees, M. J. & Mészáros, P. Shocked by GRB 970228: the afterglow of a cosmological fireball. *Mon. Not. R. Astron. Soc.* **288,** 51–56 (1997).
5. Zhang, B. & Mészáros, P. Gamma-ray bursts: progress, problems and prospects. *Int. J. Mod. Phys. A* **19,** 2385–2472 (2004).
6. Sari, R., Piran, T. & Narayan, R. Spectra and light curves of gamma-ray burst afterglows. *Astrophys. J.* **497,** L17–L20 (1998).
7. Vreeswijk, P. M. *et al.* VLT/UVES redshift of GRB 080319B. *GCN Circ.* **7444** (2008).
8. Castro-Tirado, A. J. *et al.* Decay of the GRB 990123 optical afterglow: implications for the fireball model. *Science* **283,** 2069–2073 (1999).
9. Akerlof, C. *et al.* Observation of contemporaneous optical radiation from a γ-ray burst. *Nature* **398,** 400–402 (1999).
10. Blake, C. H. *et al.* An infrared flash contemporaneous with the γ-rays of GRB 041219a. *Nature* **435,** 181–184 (2005).
11. Vestrand, W. T. *et al.* A link between prompt optical and prompt γ-ray emission in γ-ray bursts. *Nature* **435,** 178–180 (2005).
12. Barthelmy, S. D. *et al.* The Burst Alert Telescope (BAT) on the SWIFT Midex Mission. *Space Sci. Rev.* **120,** 143–164 (2005).
13. Racusin, J. L. *et al.* GRB 080319B: Swift detection of an intense burst with a bright optical counterpart. *GCN Circ.* **7427** (2008).
14. Golenetskii, S. *et al.* Konus-Wind observation of GRB 080319B. *GCN Circ.* **7482** (2008).
15. Aptekar, R. L. *et al.* Konus-W Gamma-Ray Burst Experiment for the GGS Wind Spacecraft. *Space Sci. Rev.* **71,** 265–272 (1995).
16. Cwiok, M. *et al.* Search for GRB related prompt optical emission and other fast varying objects with 'Pi of the Sky' detector. *Astrophys. Space Sci.* **309,** 531–535 (2007).
17. Molinari, E. *et al.* TORTOREM: Two-telescope complex for detection and investigation of optical transients. *Nuovo Cimento B* **121,** 1525–1526 (2006).
18. Pagani, C. *et al.* Swift observation of GRB 080319A. *GCN Rep.* **121** 1, (2008).
19. Zerbi, F. M. *et al.* The REM telescope: detecting the near infra-red counterparts of gamma-ray bursts and the prompt behavior of their optical continuum. *Astron. Nachr.* **322,** 275–285 (2001).
20. Roming, P. W. A. *et al.* The Swift Ultra-Violet/Optical Telescope. *Space Sci. Rev.* **120,** 95–142 (2005).
21. Burrows, D. B. *et al.* The Swift X-Ray Telescope. *Space Sci. Rev.* **120,** 164–195 (2005).
22. Bloom, J. *et al.* Observations of the naked-eye GRB 080319B: implications of nature's brightest explosion. Preprint at ⟨http://arxiv.org/abs/0803.3215⟩ (2008).
23. Kumar, P. & Panaitescu, A. What did we learn from gamma-ray burst 080319B? Preprint at ⟨http://arxiv.org/abs/0805.0144⟩ (2008).
24. Kobayashi, S. Light curves of gamma-ray burst optical flashes. *Astrophys. J.* **545,** 807–812 (2000).
25. Nakar, E. & Piran, T. Early afterglow emission from a reverse shock as a diagnostic tool for gamma-ray burst outflows. *Mon. Not. R. Astron. Soc.* **353,** 647–653 (2004).

26. Ramirez-Ruiz, E. & Fenimore, E. E. Pulse width evolution in gamma-ray bursts: evidence for internal shocks. *Astrophys. J.* **539,** 712–717 (2000).
27. Sari, R. & Piran, T. GRB 990123: The optical flash and the fireball model. *Astrophys. J.* **517,** L109–L112 (1999).
28. Mészáros, P. & Rees, M. J. GRB 990123: reverse and internal shock flashes and late afterglow behaviour. *Mon. Not. R. Astron. Soc.* **306,** 39–43 (1999).
29. Panaitescu, A. & Mészáros, P. Gamma-ray bursts from upscattered self-absorbed synchrotron emission. *Astrophys. J.* **544,** L17–L21 (2000).
30. Kumar, P. & McMahon, E. A general scheme for modelling γ-ray burst prompt emission. *Mon. Not. R. Astron. Soc.* **384,** 33–63 (2008).
31. Steinle, H. *et al.* Measurements of gamma-ray bursts with GLAST. *Chinese J. Astron. Astrophys.* **6** (Suppl. S1), 365–368 (2006).
32. Pedersen, H. *et al.* Evidence for diverse optical emission from gamma-ray burst sources. *Astrophys. J.* **496,** 311–315 (1998).
33. Frail, D. *et al.* The enigmatic radio afterglow of GRB 991216. *Astrophys. J.* **538,** L129–L132 (2000).
34. Ramirez-Ruiz, E., Celotti, A. & Rees, M. J. Events in the life of a cocoon surrounding a light, collapsar jet. *Mon. Not. R. Astron. Soc.* **337,** 1349–1356 (2002).
35. Kumar, P. & Piran, T. Energetics and luminosity function of gamma-ray bursts. *Astrophys. J.* **535,** 152–157 (2000).
36. Peng, F., Königl, A. & Granot, J. Two component jet models of gamma-ray burst sources. *Astrophys. J.* **626,** 966–977 (2005).
37. Berger, E. *et al.* A common origin for cosmic explosions inferred from calorimetry of GRB030329. *Nature* **426,** 154–157 (2003).
38. Huang, Y. F. *et al.* Rebrightening of XRF 030723: further evidence for a two-component jet in a gamma-ray burst. *Astrophys. J.* **605,** 300–306 (2004).
39. Zhang, W., Woosley, S. E. & MacFadyen, A. I. Relativistic jets in collapsars. *Astrophys. J.* **586,** 356–371 (2003).
40. Kumar, P. & Panaitescu, A. Afterglow emission from naked gamma-ray bursts. *Astrophys. J.* **541,** L51–L54 (2000).
41. Zhang, B. & Kobayashi, S. Gamma-ray burst early afterglows: reverse shock emission from an arbitrarily magnetized ejecta. *Astrophys. J.* **628,** 315–334 (2005).
42. Zhang, B., Kobayashi, S. & Mészáros, P. Gamma-ray burst early optical afterglows: implications for the initial Lorentz factor and the central engine. *Astrophys. J.* **595,** 950–954 (2003).
43. Kumar, P. & Panaitescu, A. A unified treatment of the gamma-ray burst 021211 and its afterglow. *Mon. Not. R. Astron. Soc.* **346,** 905–914 (2003).
44. Chevalier, R. A. & Li, Z. Y. Wind interaction models for gamma-ray burst afterglows: the case for two types of progenitors. *Astrophys. J.* **536,** 195–212 (2000).
45. Sari, R., Piran, T. & Halpern, J. Jets in gamma-ray bursts. *Astrophys. J.* **519,** L17–L20 (1999).
46. Cordes, J. M. & Lazio, T. J. W. NE2001. I. A new model for the galactic distribution of free electrons and its fluctuations. Preprint at ⟨http://arxiv.org/PS_cache/astro-ph/pdf/0207/0207156v3.pdf⟩ (2002).
47. Walker, M. A. Interstellar scintillation of compact extragalactic radio sources. *Mon. Not. R. Astron. Soc.* **294,** 307–311 (1998).
48. Frail, D. *et al.* Beaming in gamma-ray bursts: evidence for a standard energy reservoir. *Astrophys. J.* **562,** L55–L58 (2001).
49. Soderberg, A. *et al.* Radio detection of GRB 080319B. *GCN Circ.* **7506** (2008).
50. Band, D. *et al.* BATSE observations of gamma-ray burst spectra. I. Spectral diversity. *Astrophys. J.* **413,** 281–292 (1993).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank E. Rol for comments. This research was supported by NASA, the National Science Foundation (NSF), the Agenzia Spaziale Italiana, the Ministero dell'Università e della Ricerca (MUR), the Ministero degli Affari Esteri, the Netherlands Organization for Scientific Research (NWO), the National Science Foundation of China, the Russian Space Agency, Science and Technology and Facilities Council (STFC), the Slovenian Research Agency, the Ministry for Higher Education, Science, and Technology, Slovenia, and the Polish Ministry of Science and Higher Education.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.L.R. (racusin@astro.psu.edu).

ARTICLES

Strigolactone inhibition of shoot branching

Victoria Gomez-Roldan[1], Soraya Fermas[2], Philip B. Brewer[3], Virginie Puech-Pagès[1], Elizabeth A. Dun[3], Jean-Paul Pillot[2], Fabien Letisse[4], Radoslava Matusova[5], Saida Danoun[1], Jean-Charles Portais[4], Harro Bouwmeester[5,6], Guillaume Bécard[1], Christine A. Beveridge[3,7]*, Catherine Rameau[2]* & Soizic F. Rochange[1]*

A carotenoid-derived hormonal signal that inhibits shoot branching in plants has long escaped identification. Strigolactones are compounds thought to be derived from carotenoids and are known to trigger the germination of parasitic plant seeds and stimulate symbiotic fungi. Here we present evidence that *carotenoid cleavage dioxygenase 8* shoot branching mutants of pea are strigolactone deficient and that strigolactone application restores the wild-type branching phenotype to *ccd8* mutants. Moreover, we show that other branching mutants previously characterized as lacking a response to the branching inhibition signal also lack strigolactone response, and are not deficient in strigolactones. These responses are conserved in *Arabidopsis*. In agreement with the expected properties of the hormonal signal, exogenous strigolactone can be transported in shoots and act at low concentrations. We suggest that endogenous strigolactones or related compounds inhibit shoot branching in plants. Furthermore, *ccd8* mutants demonstrate the diverse effects of strigolactones in shoot branching, mycorrhizal symbiosis and parasitic weed interaction.

More than a decade ago, increased branching mutants in garden pea (*Pisum sativum* L.) and petunia (*Petunia hybrida*), *rms1* and *dad1* respectively, revealed that a mobile signal produced in shoot and root inhibits shoot branching[1,2]. For simplicity, we refer to this signal as SMS (shoot multiplication signal)[3]. SMS moves acropetally in shoots and inhibits lateral bud outgrowth[4]. On the basis of measurements of cytokinin, auxin and abscisic acid in *rms1* mutants, and in a similar mutant *rms5*, SMS was concluded to be a novel plant hormone[2,5,6]. *MAX3* and *MAX4* shoot branching genes in *Arabidopsis thaliana* encode carotenoid cleavage dioxygenases CCD7 and CCD8 (refs 7–9), and are orthologous to *RMS5* and to *RMS1* in pea, respectively[10,11] acting in the synthesis of SMS (refs 5, 10, 11). Although we cannot be sure of their role *in planta*, on the basis of studies in *Escherichia coli*, CCD7 and CCD8 may be involved in the sequential cleavage of β-carotene[7,8,12], indicating that SMS may be carotenoid derived. *CCD7* and *CCD8* are conserved across angiosperm species including monocotyledons[3,13,14]. Grafting studies indicate that *RMS4* in pea and *MAX2* in *Arabidopsis*, which encode F-box leucine-rich repeat proteins[11,15,16], confer response to SMS[2,17,18].

The strategy used here to identify SMS relies on the suggestion that SMS may be an identified carotenoid-derived compound that has already been described as controlling another biological process. Strigolactones are examples of such compounds.

Strigolactones are thought to be the principal plant-derived signal that promotes seed germination of the parasitic plants *Striga* and *Orobanche*[19,20]. Recent data suggest that strigolactones might also act in arbuscular mycorrhizal symbiosis. This interaction, probably the most widespread and significant symbiosis in nature[21], associates Glomeromycota fungi with the roots of most land plants[22]. Strigolactones are thought to function as an early host plant recognition signal, acting at picomolar concentrations to trigger the characteristic hyphal branching of arbuscular mycorrhizal fungi before root colonization[23,24]. They are found in root exudates of many species including dicotyledons and monocotyledons[20] and have also been detected in shoots[25,26]. The fact that they are also present in the non-mycotrophic plant *Arabidopsis*[27] suggests that they may have additional functions in plants. In 2005, one study[28] showed that plants treated with fluridone, an inhibitor of an early enzyme in the biosynthetic pathway of carotenoids, had reduced stimulation of the germination of parasitic plant seeds. That study also proposed a biochemical pathway in which β-carotene is cleaved to produce the strigolactones[28].

No strigolactone biosynthesis mutants have previously been identified. Given the known properties of strigolactones, our approach first involved findings relating to fungal symbiosis and parasitic seed germination, and then finally to shoot branching. We showed that *P. sativum ccd8* mutations affect mycorrhizal symbiosis and that this could be restored with exogenous strigolactone. We then observed depleted strigolactone content in *P. sativum ccd8* samples. Finally, we observed shoot branching inhibition by exogenous strigolactone in these mutants.

Mycorrhizae and parasitic plant responses

As CCD7 and CCD8 might be involved in β-carotene metabolism[7,8,12], the corresponding mutants are candidates for strigolactone deficiency[28]. As detection of strigolactones is not trivial, we chose to screen the mutants for arbuscular mycorrhizal fungal hyphae branching and parasitic seed germination defects. We chose garden pea because it can be infected by arbuscular mycorrhizal fungi and parasitic plants and used *rms1* (*P. sativum ccd8*) as the type-line for this phenotype. Enhanced branching of arbuscular mycorrhizal fungal hyphae has been used as a bioassay for the identification of strigolactones[23,24]. Root exudates of *ccd8* plants had significantly reduced activity in promoting fungal hyphae branching when compared with wild-type exudates (Fig. 1a). Similar fungal hyphae branching results were obtained with another arbuscular mycorrhizal fungal species, *Gigaspora gigantea*, and with root exudates of *ccd7* plants (data not shown).

Another key feature of strigolactones is that they stimulate the seed germination of parasitic plants such as *Orobanche*, a natural parasite

[1]Université de Toulouse; UPS; CNRS; Surface Cellulaire et Signalisation chez les Végétaux, 24 chemin de Borde Rouge, F-31326 Castanet-Tolosan, France. [2]Station de Génétique et d'Amélioration des Plantes, Institut J. P. Bourgin, UR254 INRA, F-78000 Versailles, France. [3]ARC Centre of Excellence for Integrative Legume Research, The University of Queensland, Brisbane 4072, Australia. [4]CNRS, UMR5504, INRA, UMR792 Ingénierie des Systèmes Biologiques et des Procédés, INSA de Toulouse, F-31400 Toulouse, France. [5]Plant Research International, PO Box 16, 6700 AA Wageningen, the Netherlands. [6]Laboratory of Plant Physiology, Wageningen University, Arboretumlaan 4, 6703 BD Wageningen, the Netherlands. [7]School of Integrative Biology, The University of Queensland, Brisbane 4072, Australia.
*These authors contributed equally to this work.

of pea. The germination of *Orobanche* seeds exposed to *ccd8* mutant exudates of pea was markedly decreased compared with wild-type exudates (Fig. 1b). Similar results were obtained for *ccd7* plants (data not shown). This demonstrates that *ccd7* and *ccd8* mutants lack host recognition signals required for this interaction. In addition, three concentrations of the synthetic strigolactone analogue GR24 (ref. 23), commonly used in strigolactone research, were tested in both bioassays (hyphal branching and *Orobanche* germination) and showed that relatively small changes in biological responses were induced by large differences in strigolactone concentration (Fig. 1a, b). The reduced activities observed in *ccd8* mutant root exudates of pea may therefore reflect a large decrease in strigolactone content.

To test whether the effects on fungal hyphae branching translate to reduced mycorrhizal colonization in *P. sativum ccd8* plants, we examined arbuscular mycorrhizal symbiosis with the symbiont *Glomus intraradices*. The extent of root colonization of these *ccd8* mutants was reduced compared to wild type (Fig. 1c). Similar results were obtained with other pea *ccd8* mutant lines, and after inoculation with another arbuscular mycorrhizal fungus, *Gigaspora rosea* (data not shown). Supplementation with GR24 restored the capacity of *ccd8* mutant plants to develop arbuscular mycorrhizal symbiosis and did not affect mycorrhization of wild-type plants (Fig. 1c). Together, these observations indicate an important role for strigolactones in the arbuscular mycorrhizal interaction.

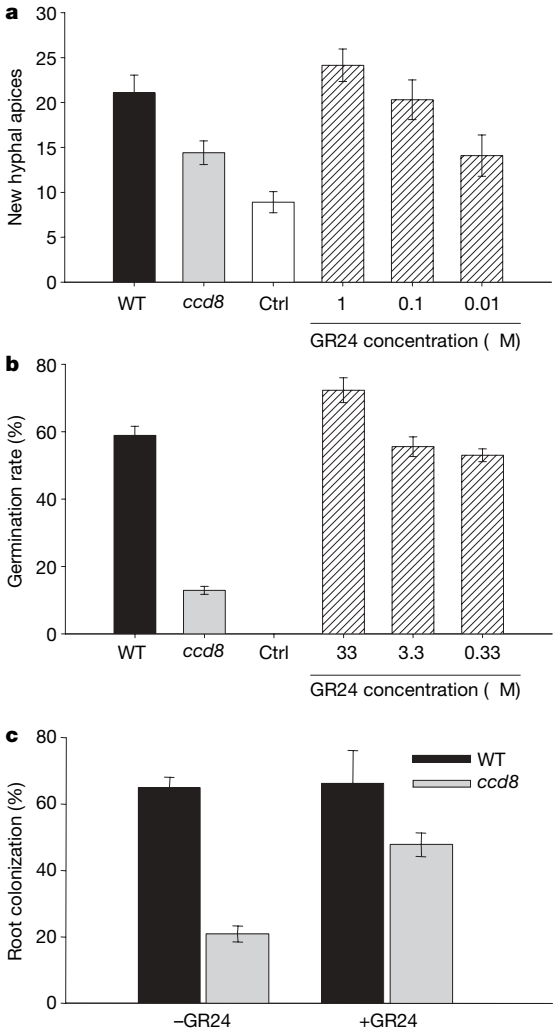


Figure 1 | Mycorrhizae and parasitic plant responses in pea *ccd8* mutants. **a**, Branching response of germinated spores of *Gigaspora rosea* to wild-type (WT) or *rms1-10* (*ccd8*) root exudate extracts, 10% acetonitrile only (Ctrl) or GR24. **b**, Germination rate of *Orobanche crenata* seeds 7 days after treatment with root exudate extracts, water (Ctrl) or GR24. **c**, Percentage of root colonization by *Glomus intraradices* of wild-type and *rms1-3* (*ccd8*) mutant plants, treated or not with 10 nM GR24. Data are means ± s.e.m. (**a**, $n > 10$; **b**, $n = 3$ with 80–100 seeds per replicate; **c**, $n = 4$).

Strigolactone analysis

The nine strigolactones identified so far share a common four-ring backbone (Fig. 2a) and differ from one another by the degree of saturation of rings A and B and the combinations of substituents that they carry[29].

Using liquid chromatography/tandem mass spectrometry (LC/MS–MS), we searched for molecules that on fragmentation yield a daughter ion at *m/z* 97 (D ring), characteristic of strigolactones, by using the precursor ion mode. This method allowed detection of two strigolactones in wild-type pea root exudate extracts (Supplementary Fig. 1). Compound 1 comprised ions at *m/z* 411 $[M + Na]^+$, *m/z* 389 $[M + H]^+$ and *m/z* 347 $[M + H - CH_2CO]^+$ (Fig. 2a), suggesting that this compound is orobanchyl acetate ($C_{21}H_{24}O_7$)[30], a major strigolactone reported recently in pea[31]. Co-chromatography of an orobanchyl acetate standard and the endogenous compound was observed using LC/MS–MS in the multiple reaction monitoring (MRM) mode (Fig. 2b). Furthermore, accurate mass MS and MS–MS data determined by ultra-performance liquid chromatography with quadruple time-of-flight mass spectrometry (UPLC/QTOFMS) for compound 1 in the sample match that of the orobanchyl acetate standard (Fig. 2c, d), and all ions are within 30 parts per million (p.p.m.) of the corresponding theoretical masses. We conclude that compound 1 is orobanchyl acetate.

The second compound was characterized by ions at *m/z* 422, 405 and 345 (Fig. 2e). In UPLC/QTOFMS, this compound showed ions at *m/z* 405.1555 and *m/z* 427.1377 (Fig. 2f). These masses were within 2 p.p.m. of the theoretical masses for the proton and sodium adducts of a molecule of elemental composition $M = C_{21}H_{24}0_8$ (404 Da). This compound could correspond to a putative strigolactone recently reported in pea root exudates with a molecular mass of 404 Da[31]. The accurate masses of the daughter ions after MS–MS (Fig. 2f) were equivalent (with a difference of one oxygen atom) to those produced by collision-induced dissociation of orobanchyl acetate[30] (Fig. 2c). The daughter ions at *m/z* 97.0285 ($[D\ ring]^+$) and 248.1058 ($[M + H - D\ ring - CH_3COOH]^+$) support the identification of this compound as an acetylated strigolactone. Although its exact structure could not be determined, it can be speculated that this compound corresponds to an acetylated strigolactone with an epoxy or hydroxyl function.

Subsequently, we analysed samples by LC/MS–MS in the MRM mode to maximize sensitivity. Characteristic transitions (precursor ion > daughter ion) for orobanchyl acetate and the second strigolactone were monitored. No other known strigolactones could be detected (data not shown). Strigolactones were analysed in the root exudates of wild-type pea, the SMS synthesis mutant *ccd8*, and in the SMS response mutant *rms4*. Both strigolactones were detected in wild-type and *rms4* exudates but were undetectable in *ccd8* exudates (Fig. 3). Analyses of root tissue extracts gave similar results (data not shown). Strigolactone deficiency in *P. sativum ccd* mutant root exudates was confirmed with independent alleles of *ccd8* and of *ccd7* (data not shown), demonstrating that the effect was due to the *ccd* mutations. The depleted strigolactone content in these *ccd8* mutant exudates (Fig. 3) combined with the reduced mycorrhizal colonization in *ccd8* plants, which can be restored by exogenous strigolactone (Fig. 1c), supports the hypothesis that CCD8 in pea controls strigolactone content and arbuscular mycorrhizal interactions.

Strigolactone application inhibits shoot branching

If SMS is a strigolactone in pea, then outgrowth of buds in *P. sativum ccd8* plants should be inhibited by an active strigolactone whereas those of *rms4* should not. By applying the synthetic strigolactone analogue GR24 directly to the axillary buds of *ccd8* and *rms4* plants

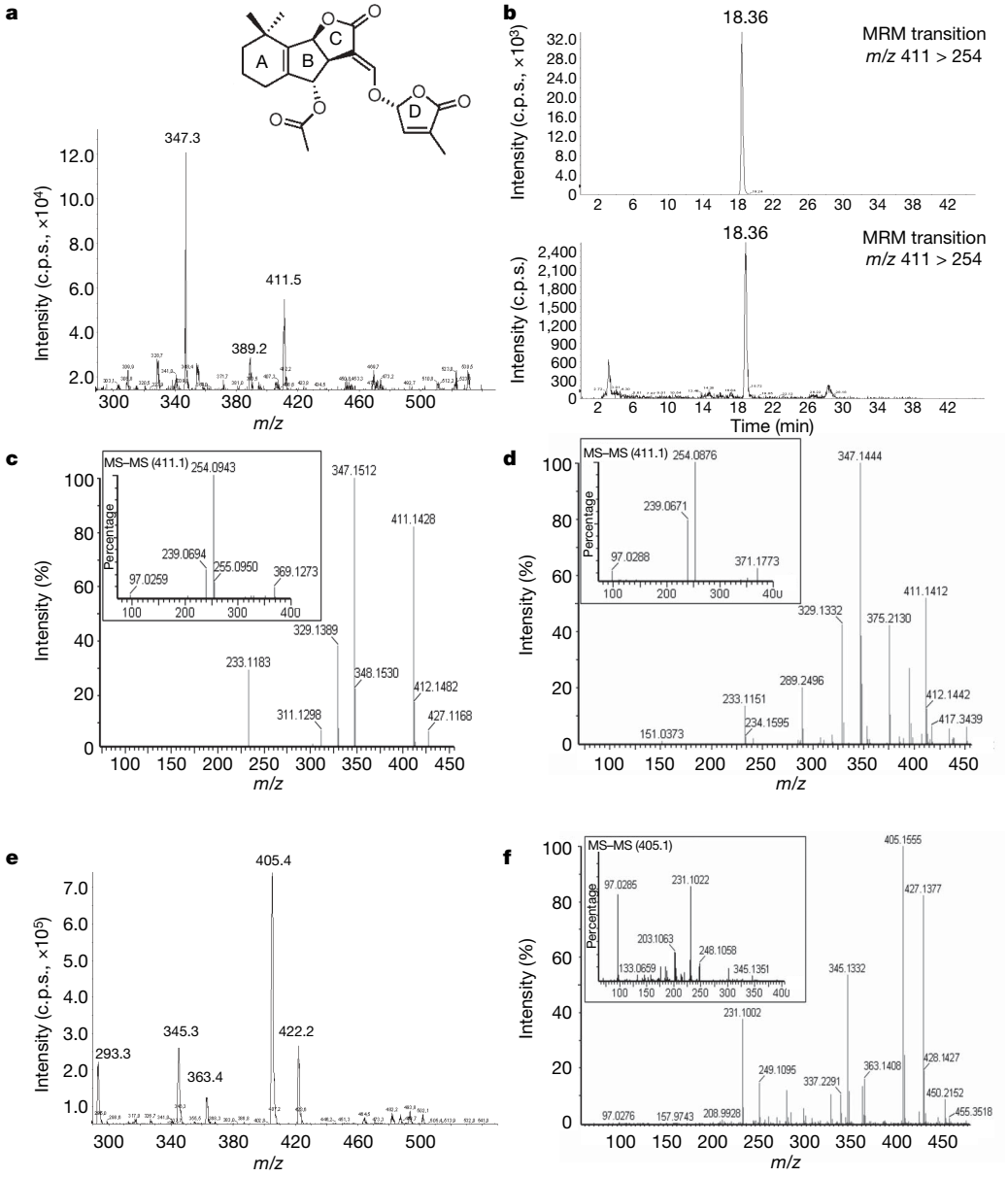


Figure 2 | Detection of two strigolactones in pea by LC/MS-MS. a, Positive ion mass spectrum of compound 1 extracted from the precursor ion contour plot, and structure of orobanchyl acetate. c.p.s., counts per second. **b**, One-channel MRM chromatogram of an orobanchyl acetate standard (top) and wild-type root exudate extract (bottom). **c**, **d**, UPLC/QTOFMS and MS–MS spectra obtained for an orobanchyl acetate standard (**c**) and compound 1 in wild-type root exudate extract (**d**). **e**, Positive ion mass spectrum of compound 2 extracted from the precursor ion contour plot. **f**, UPLC/QTOFMS and MS–MS spectra obtained for compound 2 in the wild-type root exudate extract.

we observed the expected result that the SMS-deficient *ccd8* plants responded to the treatment, whereas the SMS-insensitive *rms4* plants did not (Fig. 4a). Direct application of a greater strigolactone dose (500 nM GR24) also failed to inhibit outgrowth in *rms4* plants (data not shown). Axillary buds of *ccd8* plants were greatly inhibited by only one 10 µl application of 100 nM GR24 supplied directly to the bud. In contrast, buds of control *ccd8* plants grew 14 mm over 8 days (Fig. 4a).

As SMS is known to act as a long-distance signal and move acropetally in shoots[4] we developed a method to feed GR24 to the vascular stream of shoots (see Methods). Ten nanomolar or higher concentrations of GR24 were sufficient to inhibit bud outgrowth in *P. sativum ccd8* plants at the two nodes above the feeding site (Fig. 4b; data not shown). *ccd7* plants showed a similar response as *ccd8* plants (data not shown). As the pulse of solution supplied was diluted into the endogenous transport stream and transported widely throughout the shoot system above (dye studies; data not shown), it is probable that GR24 is highly active, consistent with the expected properties of SMS based on interstock grafting experiments[1,4,7].

To test whether other plant species are able to respond to strigolactone-mediated shoot branching inhibition, we tested the response of *Arabidopsis thaliana* branching mutants to application of GR24 to leaf axils and axillary buds before and during bolting (Fig. 5). In *A. thaliana ccd8* mutant plants, shoot branching decreased in response to GR24. In addition, the *max1* mutant of *Arabidopsis*, which acts downstream of *ccd8* in the synthesis of SMS[18], responded to GR24 (Fig. 5). Moreover, branching in the SMS response mutant *max2* (refs 15, 32) was not inhibited by GR24. These results are expected if the SMS in *Arabidopsis* is a strigolactone.

As individual buds were treated in pea, the repression by GR24 may be reinforced by dominance from non-treated buds[33]. In contrast, all *Arabidopsis* buds were treated simultaneously, possibly nullifying this effect. Alternatively, *Arabidopsis* plants may require earlier applications for a full response. The increase in branching seen in

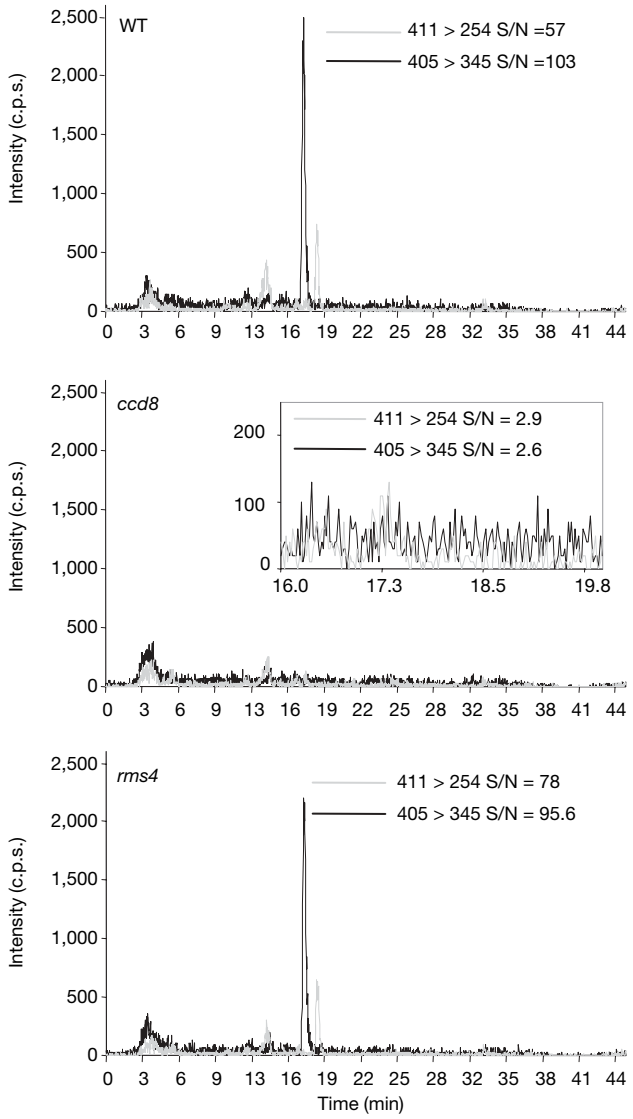


Figure 3 | *ccd8* mutant root exudates of pea are deficient in strigolactones compared to wild-type and *rms4* plants. Characteristic transitions for orobanchyl acetate (grey line) and the second strigolactone (black line) were monitored in the MRM mode. The relevant part of the *rms1-10* (*ccd8*) chromatogram is enlarged in the inset; no significant peaks are seen. Signal-to-noise (S/N) ratios are given for both peaks observed in wild-type (WT) and *rms4* plants and for signals measured at the same retention times in *ccd8* plants.

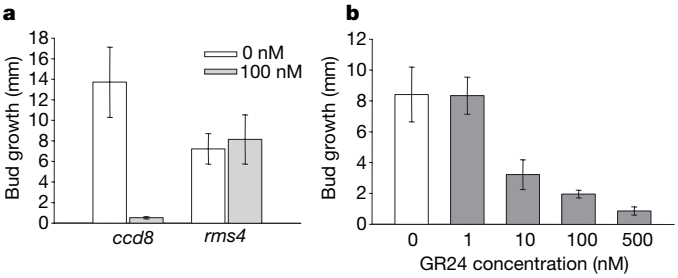


Figure 4 | GR24 inhibits bud outgrowth in pea. **a**, Axillary buds of *rms1-10* (*ccd8*) plants respond to GR24 whereas those of *rms4-3* do not. GR24 (0 or 100 nM) was applied directly to the bud at node 4. **b**, GR24 inhibits bud outgrowth of *rms1-10* plants at low concentrations and can move acropetally in shoots. GR24 and control solutions were supplied to the vascular stream between nodes 3 and 4. **a**, **b**, Bud growth was measured at node 4, 8 days (**a**) and node 5, 10 days (**b**) after treatment. Data are means ± s.e.m. (**a**, $n = 9$; **b**, $n = 8$).

max2-treated plants, but not in *rms4* plants (Figs 4a and 5), may be due to the different treatment methods or to differences in feedback regulation between the two species[3]. Indeed, the *max2*-increased branching response is not observed under all treatment conditions (data not shown).

The response of *ccd8* and *rms4* mutants to strigolactone therefore spans the classical criteria expected for hormone action. These include removal via genetic approaches and replacement by exogenous hormone treatment, as well as sensitivity, specificity and long-distance signalling. We observed reduced strigolactone content in *P. sativum ccd8* mutants compared to wild-type plants (Fig. 3), and branching is suppressed when strigolactone is supplied directly to the bud or supplied within the stem below the bud (Fig. 4). The branching inhibition response when strigolactone is supplied within the stem below a bud is consistent with the expectation that SMS moves acropetally and can act at a distance[4,34]. Moreover, the ability of GR24 to inhibit branching when added directly to buds implies that it may act at this site. Furthermore, the dose–response to strigolactone from 10 nM and above is consistent with the dose relationship and sensitivity expected for a plant hormone (Fig. 4b)[35–37]. *rms4* mutants, characterized as defective in SMS response, do not show shoot branching inhibition in response to strigolactone (Fig. 4a) and do not have reduced strigolactone content (Fig. 3). This is consistent with strigolactones acting downstream of CCD8 and upstream of RMS4 and demonstrates that the effect of strigolactones is specific to the SMS signalling pathway. Consistent with the suggested conservation of the SMS biosynthesis and signalling pathway in plants[3,13,14], strigolactone application also inhibited branching in SMS synthesis mutants (*ccd8* and *max1*) but not in the SMS response mutant (*max2*) of *Arabidopsis*. Other than the effects on shoot branching, we observed no other effects of our strigolactone treatments to shoots, again consistent with the non-pleiotropic phenotype of these mutants.

These data raise several possibilities. Either the novel hormone SMS is a strigolactone or closely related molecule, strigolactones regulate the level of SMS, or SMS regulates the level of strigolactones, which in turn inhibit branching. Many plant hormones act in tandem with other hormones. Indeed, auxin may control shoot branching by regulating SMS[38]. Nevertheless, strigolactones are now the strongest candidate for SMS and adhere to the criteria expected for a plant hormone. Future studies will focus on confirming the identity of the active molecule(s) *in planta*, and on their biosynthesis, reception and signal transduction pathways. This will require purified or synthetic strigolactones identical to the naturally occurring molecules with and without isotope labelling, and the identification of active and inactive analogues. In the absence of multiple biosynthetic and metabolic mutants on the SMS pathway, conclusive evidence for the identity

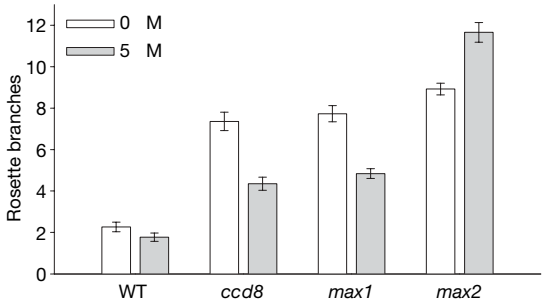


Figure 5 | GR24 inhibits bud outgrowth in *Arabidopsis*. Treatments of 0 or 5 μM GR24 were applied to the rosette axillary buds and leaf axils of wild-type, *max4-1* (*ccd8*), *max1-1* and *max2-1* plants (Columbia ecotype). Plants were treated every third day for 20 days from 23 days of age. The number of branches was counted on 48-day-old plants. Data are means ± s.e.m. ($n = 25$–41).

of SMS may involve the binding of a molecule to a putative receptor target, perhaps RMS4/MAX2, known to act in buds or the adjacent stem to inhibit branching.

Implications

This discovery provides the basis for applications in forestry, horticulture and crop science where these natural compounds, unlike other hormones such as cytokinin and auxin, may be used to specifically regulate shoot branching with minimal side effects and without the need for transgenic technology.

Our data provide the first genetic evidence for the importance of strigolactones in two very different interactions with plants—one beneficial (arbuscular mycorrhizal symbiosis) and the other detrimental (plant parasitism)—as well as evidence that the novel hormone that controls shoot branching may belong to the strigolactone family. Thus, strigolactones may act as internal as well as rhizosphere signals. Interestingly, both shoot branching and mycorrhizal colonization of plants date to the very early development of land plants. Arbuscular mycorrhizal fungi arose at a time when the land flora consisted mostly of bryophytes and it is likely that they had a crucial role in facilitating the colonization of the land by plants[39]. The emergence of axillary meristems and branching in vascular plants was critical for the development of large plants[40]. Genes probably encoding CCD7 and CCD8 occur in the moss *Physcomitrella patens*, indicating that the strigolactone biosynthetic pathway may have developed concomitantly with these two major innovations. In mycotrophic plants, strigolactone levels in root exudates are markedly affected by nutrient supply, which may relate to a putative control by plants of their degree of colonization by arbuscular mycorrhizal fungi according to their needs[25,41]. Whether plants have evolved mechanisms to control strigolactone synthesis and localization independently for the stimulation of symbiosis and the control of shoot architecture deserves further investigation.

The fact that only three genes so far, *CCD7*, *CCD8* and *MAX1*, have been identified as potentially encoding strigolactone biosynthetic enzymes is not unlike auxin and cytokinin biosynthetic pathways where genetic redundancy has greatly hampered the isolation of mutants from phenotypic screens[42,43]. Indeed, *MAX1* is represented by multiple copies in other species investigated, including pea and rice (data not shown). Strigolactone application combined with microarray studies will be useful for the isolation and characterization of branching mutants.

Notwithstanding the possibility that new roles for strigolactones are yet to be discovered, strigolactones may differ from other plant hormones by their apparent specific action in plants, as shown by the phenotype of the strigolactone-related mutants and the non-pleiotropic response to strigolactone application. Highly pleiotropic plant hormones, such as auxin[44], can mask the role of other more specific hormones. Indeed, this was the case for strigolactones, as for many years auxin, cytokinin and abscisic acid were thought to be the only hormones that control branching[2]. For shoot branching, grafting was an essential tool to demonstrate the existence of a novel long-distance signal, and to separate its role from other hormones[2,45]. This report of the inhibitory effect of a strigolactone on axillary bud outgrowth, together with the availability of genetic and biochemical tools, should allow the determination of the mode of action of strigolactones in this process and help resolve the debate on how auxin affects shoot branching[13,46]. Such progress will be an important step towards a deeper understanding of plant development and its complex regulation.

METHODS SUMMARY

For the collection of root exudates, 4-week-old plants were removed from pots, their roots were rinsed and the plants were kept for 24 h in phosphate-free Long Ashton nutrient solution (LANS)[47]. Root exudates were extracted with one volume of ethyl acetate and dried. *Gigaspora rosea* hyphal branching bioassays were performed as described[24] with partially purified root exudate samples. The number of newly formed hyphal apices was determined 48 h after treatment. *Orobanche crenata* seed germination bioassays were carried out as described[28]; seeds were treated with 50 µl exudate extracts corresponding to 4 mg root dry weight. For the determination of mycorrhizal capacity, germinated pea seeds were planted in pots containing 300 spores of *Glomus intraradices* or 30 spores of *Gigaspora rosea*. The percentage of root length colonized by the fungus, that is, showing arbuscules, vesicles or both, was determined after 6 weeks in culture by the gridline intersection method[48], after staining with Schaeffer black ink[49]. Strigolactone detection was performed using a 4000 Q Trap mass spectrometer coupled to a high performance liquid chromatography (HPLC) system. Accurate masses were acquired with a UPLC/QTOF device. For shoot branching in pea, GR24 treatments were performed on 10-day-old plants. Axillary buds at node 4 were treated with 10 µl of 0 or 100 nM GR24. For vascular supply, a thread submerged in GR24 solution was passed through the stem between nodes 3 and 4 using a needle; the outgrowth of the bud at node 5 was measured. *Arabidopsis* plants were treated every third day for 20 days from 23 days of age (pre-bolting). Treatments were applied to the axillary buds or leaf axils with 50 µl per plant of 0 or 5 µM GR24. The number of rosette branches (≥5 mm) was counted on 48-day-old plants.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 17 June; accepted 18 July 2008.
Published online 10 August 2008.

1. Napoli, C. Highly branched phenotype of the petunia *dad1–1* mutant is reversed by grafting. *Plant Physiol.* **111,** 27–37 (1996).
2. Beveridge, C. A., Symons, G. M., Murfet, I. C., Ross, J. J. & Rameau, C. The *rms1* mutant of pea has elevated indole-3-acetic acid levels and reduced root-sap zeatin riboside content but increased branching controlled by graft transmissible signal(s). *Plant Physiol.* **115,** 1251–1258 (1997).
3. Beveridge, C. A. Axillary bud outgrowth: sending a message. *Curr. Opin. Plant Biol.* **9,** 35–40 (2006).
4. Foo, E., Turnbull, C. G. & Beveridge, C. A. Long-distance signaling and the control of branching in the *rms1* mutant of pea. *Plant Physiol.* **126,** 203–209 (2001).
5. Morris, S. E., Turnbull, C. G., Murfet, I. C. & Beveridge, C. A. Mutational analysis of branching in pea. Evidence that *Rms1* and *Rms5* regulate the same novel signal. *Plant Physiol.* **126,** 1205–1213 (2001).
6. Dodd, I. C., Ferguson, B. J. & Beveridge, C. A. Apical wilting and petiole xylem vessel diameter of the *rms2* branching mutant of pea are shoot controlled and independent of a long-distance signal regulating branching. *Plant Cell Physiol.* **49,** 791–800 (2008).
7. Booker, J. *et al.* MAX3/CCD7 is a carotenoid cleavage dioxygenase required for the synthesis of a novel plant signaling molecule. *Curr. Biol.* **14,** 1232–1238 (2004).
8. Auldridge, M. E. *et al.* Characterization of three members of the *Arabidopsis* carotenoid cleavage dioxygenase family demonstrates the divergent roles of this multifunctional enzyme family. *Plant J.* **45,** 982–993 (2006).
9. Bouvier, F., Isner, J. C., Dogbo, O. & Camara, B. Oxidative tailoring of carotenoids: a prospect towards novel functions in plants. *Trends Plant Sci.* **10,** 187–194 (2005).
10. Sorefan, K. *et al. MAX4* and *RMS1* are orthologous dioxygenase-like genes that regulate shoot branching in *Arabidopsis* and pea. *Genes Dev.* **17,** 1469–1474 (2003).
11. Johnson, X. *et al.* Branching genes are conserved across species. Genes controlling a novel signal in pea are coregulated by other long-distance signals. *Plant Physiol.* **142,** 1014–1026 (2006).
12. Schwartz, S. H., Qin, X. & Loewen, M. C. The biochemical characterization of two carotenoid cleavage enzymes from *Arabidopsis* indicates that a carotenoid-derived compound inhibits lateral branching. *J. Biol. Chem.* **279,** 46940–46945 (2004).
13. Mouchel, C. F. & Leyser, O. Novel phytohormones involved in long-range signaling. *Curr. Opin. Plant Biol.* **10,** 473–476 (2007).
14. Doust, A. N. Grass architecture: genetic and environmental control of branching. *Curr. Opin. Plant Biol.* **10,** 21–25 (2007).
15. Stirnberg, P., van De Sande, K. & Leyser, H. M. *MAX1* and *MAX2* control shoot lateral branching in *Arabidopsis*. *Development* **129,** 1131–1141 (2002).
16. Ishikawa, S. *et al.* Suppression of tiller bud activity in tillering dwarf mutants of rice. *Plant Cell Physiol.* **46,** 79–86 (2005).
17. Beveridge, C. A., Ross, J. J. & Murfet, I. C. Branching in pea (action of genes *Rms3* and *Rms4*). *Plant Physiol.* **110,** 859–865 (1996).
18. Booker, J. *et al. MAX1* encodes a cytochrome P450 family member that acts downstream of *MAX3/4* to produce a carotenoid-derived branch-inhibiting hormone. *Dev. Cell* **8,** 443–449 (2005).
19. Cook, C. E. *et al.* Germination stimulants. 2. The structure of strigol — a potent seed germination stimulant for witchweed (*Striga lutea* Tour.). *J. Am. Chem. Soc.* **94,** 6198–6199 (1972).
20. Bouwmeester, H. J., Roux, C., Lopez-Raez, J. A. & Becard, G. Rhizosphere communication of plants, parasitic plants and AM fungi. *Trends Plant Sci.* **12,** 224–230 (2007).
21. Brachmann, A. & Parniske, M. The most widespread symbiosis on Earth. *PLoS Biol.* **4,** e239 (2006).

22. Smith, S. E. & Read, D. J. *Mycorrhizal Symbiosis* (Academic Press, 1997).
23. Akiyama, K., Matsuzaki, K. & Hayashi, H. Plant sesquiterpenes induce hyphal branching in arbuscular mycorrhizal fungi. *Nature* **435,** 824–827 (2005).
24. Besserer, A. *et al.* Strigolactones stimulate arbuscular mycorrhizal fungi by activating mitochondria. *PLoS Biol.* **4,** e226 (2006).
25. Yoneyama, K., Yoneyama, K., Takeuchi, Y. & Sekimoto, H. Phosphorus deficiency in red clover promotes exudation of orobanchol, the signal for mycorrhizal symbionts and germination stimulant for root parasites. *Planta* **225,** 1031–1038 (2007).
26. Yoneyama, K. *et al.* Nitrogen deficiency as well as phosphorus deficiency in sorghum promotes the production and exudation of 5-deoxystrigol, the host recognition signal for arbuscular mycorrhizal fungi and root parasites. *Planta* **227,** 125–132 (2007).
27. Goldwasser, Y., Yoneyama, K., Xie, X. & Yoneyama, K. Production of strigolactones by *Arabidopsis thaliana* responsible for *Orobanche aegyptiaca* seed germination. *Plant Growth Regul.* **55,** 21–28 (2008).
28. Matusova, R. *et al.* The strigolactone germination stimulants of the plant-parasitic *Striga* and *Orobanche* spp. are derived from the carotenoid pathway. *Plant Physiol.* **139,** 920–934 (2005).
29. Xie, X. *et al.* 2′-epi-orobanchol and solanacol, two unique strigolactones, germination stimulants for root parasitic weeds, produced by tobacco. *J. Agric. Food Chem.* **55,** 8067–8072 (2007).
30. Xie, X. *et al.* Isolation and identification of alectrol as (+)-orobanchyl acetate, a germination stimulant for root parasitic plants. *Phytochemistry* **69,** 427–431 (2008).
31. Yoneyama, K. *et al.* Strigolactones, host recognition signals for root parasitic plants and arbuscular mycorrhizal fungi, from Fabaceae plants. *New Phytol.* **179,** 484–494 (2008).
32. Stirnberg, P., Furner, I. J. & Leyser, H. M. O. MAX2 participates in an SCF complex which acts locally at the node to suppress shoot branching. *Plant J.* **50,** 80–94 (2007).
33. Li, C. J. & Bangerth, F. Autoinhibition of indoleacetic acid transport in the shoots of two-branched pea (*Pisum sativum*) plants and its relationship to correlative dominance. *Physiol. Plant.* **106,** 415–420 (1999).
34. Turnbull, C. G., Booker, J. P. & Leyser, H. M. Micrografting techniques for testing long-distance signalling in *Arabidopsis*. *Plant J.* **32,** 255–262 (2002).
35. Cowling, R. J., Kamiya, Y., Seto, H. & Harberd, N. P. Gibberellin dose-response regulation of *GA4* gene transcript levels in *Arabidopsis*. *Plant Physiol.* **117,** 1195–1203 (1998).
36. Booker, J., Chatfield, S. & Leyser, O. Auxin acts in xylem-associated or medullary cells to mediate apical dominance. *Plant Cell* **15,** 495–507 (2003).
37. Nemhauser, J. L., Mockler, T. C. & Chory, J. Interdependency of brassinosteroid and auxin signaling in *Arabidopsis*. *PLoS Biol.* **2,** e258 (2004).
38. Foo, E. *et al.* The branching gene *RAMOSUS1* mediates interactions among two novel signals and auxin in pea. *Plant Cell* **17,** 464–474 (2005).
39. Remy, W., Taylor, T. N., Hass, H. & Kerp, H. Four hundred-million-year-old vesicular arbuscular mycorrhizae. *Proc. Natl Acad. Sci. USA* **91,** 11841–11843 (1994).
40. Kenrick, P. & Crane, P. R. The origin and early evolution of plants on land. *Nature* **389,** 33–39 (1997).
41. Lopez-Raez, J. A. *et al.* Tomato strigolactones are derived from carotenoids and their biosynthesis is promoted by phosphate starvation. *New Phytol.* **178,** 863–874 (2008).
42. Strader, L. C. & Bartel, B. A new path to auxin. *Nature Chem. Biol.* **4,** 337–339 (2008).
43. Miyawaki, K. *et al.* Roles of *Arabidopsis* ATP/ADP isopentenyltransferases and tRNA isopentenyltransferases in cytokinin biosynthesis. *Proc. Natl Acad. Sci. USA* **103,** 16598–16603 (2006).
44. Berleth, T., Krogan, N. T. & Scarpella, E. Auxin signals-turning genes on and turning cells around. *Curr. Opin. Plant Biol.* **7,** 553–563 (2004).
45. Beveridge, C. A., Symons, G. M. & Turnbull, C. G. N. Auxin inhibition of decapitation-induced branching is dependent on graft-transmissible signals regulated by genes *Rms1* and *Rms2*. *Plant Physiol.* **123,** 689–697 (2000).
46. Dun, E. A., Ferguson, B. J. & Beveridge, C. A. Apical dominance and shoot branching. Divergent opinions or divergent mechanisms? *Plant Physiol.* **142,** 812–819 (2006).
47. Hewitt, E. J. *Sand and Water Culture: Methods Used in the Study of Plant Nutrition* 2nd edn (London and Reading: Commonwealth Agricultural Bureau, The Eastern Press, 1966).
48. Giovannetti, M. & Mosse, B. An evaluation of techniques for measuring vesicular-arbuscular infection in roots. *New Phytol.* **84,** 489–500 (1980).
49. Vierheilig, H., Coughlan, A. P., Wyss, U. & Piche, Y. Ink and vinegar, a simple staining technique for arbuscular-mycorrhizal fungi. *Appl. Environ. Microbiol.* **64,** 5004–5007 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements V.G.-R. was supported by the French Ministry of Research and Higher Education. S.F. was supported by a grant from ANR. H. B. and R. M. were supported by a grant from The Netherlands Organisation for Scientific Research (NWO; VICI-grant). The authors are grateful to A. Marion-Poll for discussions, H. M. O. Leyser for supply of the *Arabidopsis max4* seed, K. Yoneyama for the gift of orobanchyl acetate, and D. M. Joel for providing *O. crenata* seeds. The UPLC/QTOF mass spectrometer was made available to us by the Institut des Technologies Avancées du Vivant (Toulouse, France). We thank K. Condon for plant husbandry, the ARC Centre of Excellence for Integrative Legume Research and the European Union FP6 Grain Legumes Integrated Project for financial support.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.R. (rameau@versailles.inra.fr) or G.B. (becard@scsv.ups-tlse.fr).

METHODS

Plant materials and growth conditions. Pea mutants are described in refs 50, 51. Before root exudate collection, plants were grown in autoclaved sand:sepiolite (2:1 v:v) potting mix and fertilized weekly with LANS, then watered with phosphate-free LANS for 2 days. Plants grown for the determination of mycorrhizal activity were grown in an autoclaved sand:sepiolite (2:1 v:v) potting mix at 24 °C with a 16 h photoperiod at 120 μmol m^{-2} s^{-1}. They were fertilized three times a week with a modified LANS containing 15 μM sodium phosphate, with 0 nM or 10 nM GR24. Water was supplied as needed on other days.

For shoot branching bioassays, two pea plants were grown per two-litre pot in potting mix containing peat with clay pellets in a heated glasshouse (mean 15 °C night/22 °C day) with the natural photoperiod extended to 16 h with sodium lamps (or supplemented during the day when necessary). Nodes were numbered acropetally from the first scale leaf (node 1). Lateral branches were removed from nodes 1 and 2 to encourage the outgrowth of axillary buds at nodes above. *Arabidopsis* plants were grown in a mixture of 60% University of California potting mix type C and 40% vermiculite (v/v). After sowing in punnets, seeds were vernalized by incubation at 4 °C for 2 days, and then transferred to a growth room at 21 ± 3 °C under a 16 h photoperiod with a light intensity of 50 μE m^{-2} s^{-1} provided by white fluorescent tubes.

Strigolactone treatments. For direct application on pea buds, GR24 was supplied in a solution of polyethylene glycol (PEG) 1450 (4%), ethanol (25%) and acetone (0.5%). For vascular supply, GR24 was diluted in water to the concentrations 1 nM, 10 nM, 100 nM and 500 nM. For *Arabidopsis*, GR24 was supplied in Tween20 (0.1%) and acetone (0.5%). For all treatments, control plants were treated with the same solutions without GR24.

Preparation of root exudates. For the *Gigaspora rosea* hyphal branching bioassay, exudates produced by the equivalent of 120 mg root dry weight and resuspended in 20% acetonitrile were fractionated on a 1 ml C18 Mega Bond Elut column (Varian) using an acetonitrile/water gradient (20% to 100% v:v). Fractions were dried and resuspended in 1 ml acetonitrile:water (10% v:v). Fractions corresponding to 50% and 70% acetonitrile, containing all the detectable activity, were pooled for the bioassays.

Chromatography and mass spectrometry. Root exudate extracts were dissolved in acetonitrile:water (1:1 v:v) and cleared through a 0.2 μm nylon filter. Strigolactone detection was performed using a 4000 Q Trap mass spectrometer (Applied Biosystems), with a Turbo V ESI source in the positive mode, coupled to an Agilent 1100 series HPLC system (Agilent Technologies). A C18 column (5 μm, 2.1 × 250 mm, PepMap, Dionex) was used for chromatographic separation. Solutions of formic acid:water ($1:10^3$ v:v; A) and formic acid:acetonitrile ($1:10^3$ v:v; B) were pumped at 0.2 ml min^{-1}. The gradient was: 50% B for 5 min, 50%–70% B in 5 min, 70% B for 10 min, 70%–100% B in 10 min and 100% B for 20 min.

Synthetic GR24 and sorgolactone standards (Chiralix) were used to optimize mass spectrometry parameters to the following values: nebulizer gas flow 100 l h^{-1}, desolvation gas flow 500 l h^{-1}, capillary voltage 4,000 V, source temperature 250 °C; collision gas: nitrogen, collision energy 21 V. Ten-microlitre samples produced by the equivalent of 900 mg (precursor ion mode) or 150 mg (MRM mode) root dry weight were analysed. The MRM mode was used to detect with high sensitivity the two strigolactones seen in the precursor ion mode, and also to search for previously known strigolactones. Transitions monitored in the MRM mode were the following**:** orobanchol and isomers, *m/z* 347 > 250 and 369 > 272; 5-deoxy-strigol, *m/z* 331 > 234 and 353 > 256; sorgolactone, *m/z* 317 > 220 and 339 > 242; solanacol, *m/z* 343 > 246 and 365 > 268; didehydro-orobanchol (strigol), *m/z* 367 > 270; orobanchyl acetate, *m/z* 389 > 233 and 411 > 254; and second pea strigolactone, *m/z* 405 > 97, 405 > 345, 405 > 248 and 427 > 270. Data were analysed with Analyst 1.4.1 software (Applied Biosystems).

Accurate masses were acquired with a QTOF Premier device (Waters) with an ESI source in the positive mode coupled to a UPLC system (Acquity UPLC, Waters). Chromatographic separation was performed on a BEH C18 column (1.7 μm, 2.1 × 100 mm, Acquity ULPC, Waters) applying a gradient equivalent to that used for HPLC for a duration of 9 min at a flow rate of 0.6 ml min^{-1}. Parameters were the following: nebulization gas 750 l h^{-1} at 450 °C, cone gas 30 l h^{-1}, source temperature 150 °C, capillary voltage 3 kV, cone voltage 25 V, MCP detector voltage 1,650 V; collision gas: argon, scan range 80–1,000 *m/z*, data collection in centroid mode. Data were acquired using the lock spray for accurate mass using leucine-enkephalin as the lock mass, and analysed with Micromass MassLynx application manager.

50. Arumingtyas, E. L., Floyd, R. S., Gregory, M. J. & Murfet, I. C. Branching in *Pisum*: inheritance and allelism tests with 17 *ramosus* mutants. *Pisum Genet.* **24**, 17–31 (1992).
51. Symons, G. M. & Murfet, I. C. Inheritance and allelism tests on six further branching mutants in pea. *Pisum Genet.* **29**, 1–6 (1997).

ARTICLES

Inhibition of shoot branching by new terpenoid plant hormones

Mikihisa Umehara[1], Atsushi Hanada[1], Satoko Yoshida[1], Kohki Akiyama[2], Tomotsugu Arite[3], Noriko Takeda-Kamiya[1], Hiroshi Magome[1], Yuji Kamiya[1], Ken Shirasu[1], Koichi Yoneyama[4], Junko Kyozuka[3] & Shinjiro Yamaguchi[1]

Shoot branching is a major determinant of plant architecture and is highly regulated by endogenous and environmental cues. Two classes of hormones, auxin and cytokinin, have long been known to have an important involvement in controlling shoot branching. Previous studies using a series of mutants with enhanced shoot branching suggested the existence of a third class of hormone(s) that is derived from carotenoids, but its chemical identity has been unknown. Here we show that levels of strigolactones, a group of terpenoid lactones, are significantly reduced in some of the branching mutants. Furthermore, application of strigolactones inhibits shoot branching in these mutants. Strigolactones were previously found in root exudates acting as communication chemicals with parasitic weeds and symbiotic arbuscular mycorrhizal fungi. Thus, we propose that strigolactones act as a new hormone class—or their biosynthetic precursors—in regulating above-ground plant architecture, and also have a function in underground communication with other neighbouring organisms.

Shoot branching involves the formation of axillary buds in the axil of leaves and subsequent outgrowth of the buds. Previous studies have suggested the involvement of a novel, as yet unidentified, hormone in inhibiting the outgrowth of axillary buds, using a series of recessive mutants that show enhanced shoot branching. These mutants include *ramosus* (*rms*) of pea (*Pisum sativum*)[1–4], *more axillary growth* (*max*) of *Arabidopsis*[5–9], *decreased apical dominance* (*dad*) of petunia (*Petunia hybrida*)[10,11] and *dwarf* (*d*) or *high-tillering dwarf* (*htd*) of rice (*Oryza sativa*)[12–14]. Reciprocal grafting experiments, double mutant analysis and cloning of these genetic loci suggested that the novel hormone is biosynthesized from carotenoids and moves acropetally to inhibit axillary bud outgrowth[15]. In the proposed biosynthesis pathway, *MAX3*, *RMS5* and *HTD1/D17* encode CAROTENOID CLEAVAGE DIOXYGENASE 7 (CCD7)[4,7,13], whereas *MAX4*, *RMS1*, *D10* and *DAD1* encode another subclass of CCDs designated as CCD8 (refs 6, 10, 14) (Fig. 1a). CCD7 and CCD8 might catalyse sequential carotenoid cleavage reactions, although their endogenous substrates and exact enzymatic function in plants have not been conclusive[7,16,17]. MAX1 is a cytochrome P450 monooxygenase presumably involved in a later biosynthetic step[8] (Fig. 1a). Unlike the biosynthetic mutants, the branching phenotype of the *max2*, *rms4* and *dad2* mutants is not rescued by grafting onto a wild-type rootstock, suggesting that they are insensitive to the branch-inhibiting hormone[2,8,11]. MAX2, RMS4 and D3 are orthologous members of the F-box leucine-rich repeat (LRR) protein family[4,5,12] (Fig. 1a), which probably act as the substrate recognition subunit of SCF ubiquitin E3 ligase for proteasome-mediated proteolysis[18]. The predicted biochemical function of MAX2, RMS4 and D3 is consistent with their role in signal transduction of the novel hormone.

Strigolactones are a group of terpenoid lactones (Fig. 1b) which have been found in root exudates of diverse plant species and were initially characterized as seed germination stimulants of root parasitic plants such as *Striga* and *Orobanche* species[19–21]. More recently, strigolactones were shown to act as root-derived signals for symbiotic interaction with arbuscular mycorrhizal fungi[22], which facilitate the uptake of soil nutrients by plants. This symbiosis is observed in more than 80% of terrestrial plants, coinciding with the wide distribution of this class of terpenes. Strigolactones may have additional unidentified function(s) in plants, because they induce seed germination of non-parasitic plants as well[23,24] and are also produced by non-hosts of arbuscular mycorrhizal fungi, including *Arabidopsis*[25,26]. Little is known about the biosynthesis of strigolactones. Recent work has indicated that the ABC part (Fig. 1b) is derived from carotenoids, presumably by means of the formation of oxidatively cleaved product(s)[20,27,28]. Taken together, current lines of evidence suggest that strigolactone biosynthesis involves a (epoxy)carotenoid cleavage enzyme conserved across diverse plant species. Although CCD7 and CCD8, encoded by the *MAX/RMS/DAD/D* loci, fulfil these criteria[29], their role in strigolactone biosynthesis had not been examined. Therefore, we set out to examine whether the carotenoid-derived branching inhibitor shares its biosynthetic pathway with strigolactones using rice *d* mutants.

Strigolactone levels in rice *d* mutants

To explore the potential role of D10 (CCD8) and D17 (CCD7) in strigolactone biosynthesis in rice, we analysed strigolactones in root exudates of wild-type and *d* mutants (Supplementary Fig. 1a) by liquid chromatography-quadruple/time-of-flight tandem mass spectrometry (LC/MS–MS). Because our survey of known strigolactones in hydroponic culture media of rice seedlings (cv. Shiokari) identified 2′-*epi*-5-deoxystrigol (*epi*-5DS), we synthesized deuterium-labelled *epi*-5DS (Supplementary Fig. 2) and used it as an internal standard for quantification on LC/MS–MS. We selected $[M + H]^+$ (*m/z* 332 and 331 for d_1-*epi*-5DS and cold *epi*-5DS, respectively) as parent ions on quadruple mass spectrometry and detected $[M + H - 115]^+$ (*m/z* 217.1 and 216.1 for d_1-*epi*-5DS and cold *epi*-5DS, respectively) as fragment ions on time-of-flight mass spectrometry after collision-induced dissociation for quantification (Fig. 2a, b). Full-scan spectra of fragment ions

[1]RIKEN Plant Science Center, Tsurumi, Yokohama 230-0045, Japan. [2]Graduate School of Life and Environmental Sciences, Osaka Prefecture University, 1-1 Gakuencho, Naka-ku, Sakai, Osaka 599-8531, Japan. [3]Graduate School of Agricultural and Life Sciences, The University of Tokyo, Yayoi, Bunkyo, Tokyo 113-8652, Japan. [4]Weed Science Center, Utsunomiya University, Utsunomiya 321-8505, Japan.

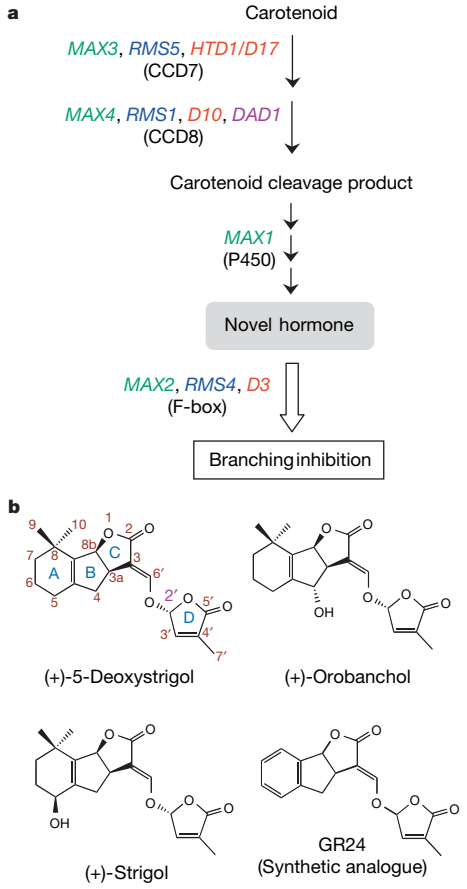


Figure 1 | The novel branching inhibitor pathway (a) and chemical structures of representative strigolactones (b).

confirmed the identity of these compounds (Fig. 2c). As observed for strigolactones in other species[28,30,31], the levels of *epi*-5DS in root exudates of wild-type seedlings were elevated when inorganic phosphate was depleted in the media (Fig. 2d). However, *epi*-5DS was nearly undetectable in exudates of *d10-1* and *d17-1* mutants, regardless of the nutrient conditions (Fig. 2d). Reduced levels of another strigolactone species (2′-*epi*-orobanchol or its isomer) in root exudates were also evident for the *d10-2* allele on the Nipponbare background (Supplementary Fig. 3). To determine whether the production of *epi*-5DS was decreased or only the secretion from roots was defective in these mutants, we quantified endogenous *epi*-5DS in roots. We found that endogenous levels of *epi*-5DS were also decreased in *d10-1* and *d17-1* seedlings relative to the wild-type control (Fig. 2e). These results demonstrate that both D10 (CCD8) and D17 (CCD7) are required for the production of normal levels of strigolactones in rice seedlings.

In contrast to the *d10-1* and *d17-1* mutants, *d3-1* seedlings accumulated higher levels of *epi*-5DS both in culture media and in roots than did wild-type plants under inorganic phosphate deficiency (Fig. 2d, e). These results are correlated with the upregulation of *D10* (*CCD8*) transcript levels in *d3-1* and other tillering *d* mutants[14], and further support the idea that D10 (CCD8) participates in strigolactone biosynthesis. Similar transcriptional regulation of *RMS1* (*CCD8*) was also found in the *rms4* mutant of pea, probably through a feedback inhibition mechanism in the branching inhibitor pathway[4]. The elevated strigolactone production in the *d3* mutant suggests that the decreased strigolactone levels in the *d10* and *d17* mutants are attributable to a direct blockage of the biosynthesis pathway, rather than a secondary consequence of the decreased branching inhibitor activity, because in the latter case, strigolactone levels would be reduced also in the *d3* mutant.

Pre-conditioned seeds of the parasitic plant *Striga hermonthica* require germination stimulants, including strigolactones, released from the host roots to complete germination. We used a highly sensitive germination assay using *S. hermonthica* seeds to estimate strigolactone concentrations in root exudates of *d* mutants[27,32]. In agreement with the LC/MS–MS data, the culture media of *d10-1* and *d17-1* seedlings contained weaker germination-stimulating activity than did those of wild-type plants (Fig. 2f). By contrast, *d3-1* root exudates exhibited stronger germination-stimulating activity than the wild-type control. The reduced germination-stimulating activity in *d10-1* root exudates is not due to increased germination inhibitors, but to decreased germination stimulants, because the addition of *d10-1* exudates did not inhibit germination induced by (+)-strigol (Fig. 2f). These results indicate that overall strigolactone levels released from roots are decreased in the *d10-1* and *d17-1* mutants.

Strigolactones inhibit tillering in rice

To investigate further the relationships between the D10/D17-derived branching inhibitor and strigolactones, we examined the effect of strigolactone treatment on rice *d* mutants. We developed a hydroponic culture system using rice seedlings, where we observed outgrowth of first and second tiller (axillary) buds in the *d* mutants, but not in the wild type. An application of GR24 (a strigolactone analogue; Fig. 1b) to the media inhibited tiller bud outgrowth of 2-week-old *d10-1* and *d17-1* seedlings in a dose-dependent manner (Fig. 3a, b). The inhibitory effect was detectable in response to as low as 10 nM GR24, and tiller bud outgrowth was nearly fully inhibited at 1 μM GR24. In contrast to *d10-1* and *d17-1*, the *d3-1* mutant, defective in a probable signalling component (Fig. 1a), was insensitive to this chemical. No morphological abnormalities were evident in wild-type seedlings after GR24 treatment. Similar effects were observed when we used naturally occurring strigolactones, (+)-strigol and (+)-5DS, as well (Figs 1b and 3c, d). The insensitivity of the *d3-1* mutant to strigolactones indicates that their inhibitory effects on tiller bud outgrowth are specific to the proposed branching inhibitor pathway. These results illustrate that strigolactones or downstream metabolites act as the novel branching inhibitor. The tillering dwarf phenotype of the *d* mutants is more drastic in appearance at a later stage[12,33]. We found that the branching phenotype as well as the plant height of 6-week-old *d10-1* mutants were complemented by including 2 μM GR24 in the culture media, whereas no visible effect of this chemical was recognizable in *d3-1* mutant plants (Fig. 3e–g). These results confirm the role of strigolactones in inhibiting tiller bud outgrowth in the branching inhibitor pathway in rice.

In many cases, hormonal responses are dose-dependent within a certain range and both hormone-deficiency and hormone-excess phenotypes are observed. We next examined the effect of a high dose of GR24 on tillering of wild-type seedlings. We found that tiller outgrowth was severely inhibited when 10 μM GR24 was supplemented to the culture media, without affecting the growth of main leaves (Supplementary Fig. 4). These observations further support the role of strigolactones in inhibiting axillary bud outgrowth and suggest the potential usefulness of strigolactones as plant growth regulators that specifically inhibit branching.

As mentioned above, *D10* transcript levels were previously shown to be elevated in the *d3-1* and *d10-1* mutants, suggesting a negative feedback control in the branch inhibitor pathway[14]. Our quantitative polymerase chain reaction with reverse transcription (qRT–PCR) analysis revealed that GR24 treatment decreased *D10* transcript levels in *d10-1* and wild-type seedlings, but not in the *d3-1* mutant (Fig. 3h). These results, together with the elevated strigolactone production in the *d3-1* mutant (Fig. 2), indicate that endogenous strigolactone levels are under homeostatic control by means of the D3-dependent signalling pathway and further support the idea that strigolactones (or downstream metabolites) act as the branching inhibitors in rice.

Effect of strigolactones in *Arabidopsis*

To determine whether strigolactones participate in the branching inhibitor pathway in *Arabidopsis*, we examined the effect of GR24

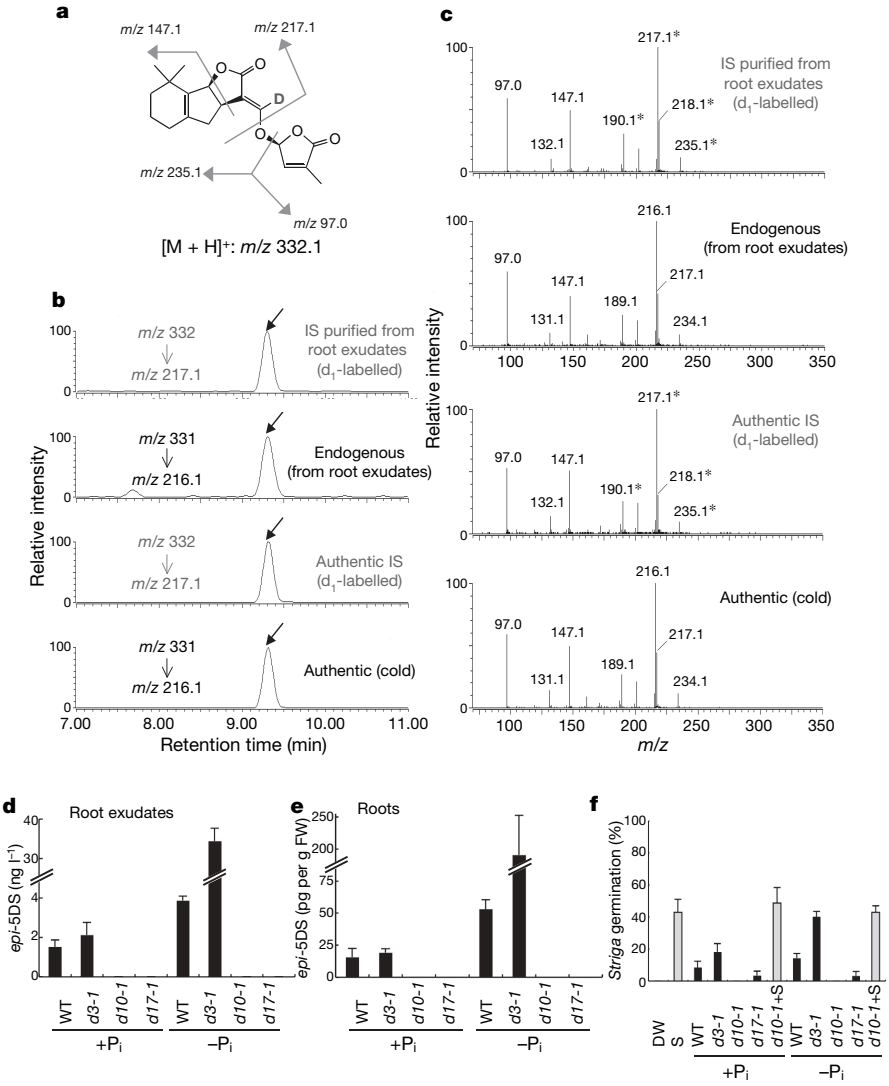


Figure 2 | Strigolactone analysis in rice seedlings. a, Predicted major fragmentation patterns of d_1-*epi*-5DS on LC/MS–MS. **b**, Selected reaction monitoring for d_1-*epi*-5DS (internal standard, IS) or *epi*-5DS. **c**, Full-scan spectra of fragment ions. Asterisks indicate deuterium-labelled ions. **d**, **e**, LC/MS–MS analysis of *epi*-5DS levels in wild type (WT) and *d* mutants in the presence ($+P_i$) or absence ($-P_i$) of inorganic phosphate (mean + s.d., $n = 3$). FW, fresh weight. **f**, Estimation of germination stimulant levels in culture media using *Striga* seeds (means and s.d., $n = 3$). DW, distilled water; S, (+)-strigol (10^{-7} M); *d10-1*+S, co-incubation with (+)-strigol (10^{-7} M) and *d10-1* culture media.

on the branching phenotype of *max* mutants (Supplementary Fig. 1b). The *MAX* genes are required for selective repression of axillary shoots and *max* mutants exhibit bushier shoots than do wild-type plants[5,9]. Our data show that the enhanced branching phenotype of *max3* and *max4* mutants (defective in CCD7 and CCD8, respectively; Fig. 1a) is rescued by supplementing the hydroponic culture media with 5 μM GR24, whereas *max2* mutants are insensitive to GR24 treatment (Fig. 4a, b). Next, we estimated the levels of strigolactones in root exudates of *max* mutants by determining germination-stimulating activity using *S. hermonthica* seeds. In root exudates from *max3* and *max4* seedlings, the levels of germination stimulants are significantly lower than those from wild-type seedlings. By contrast, the *max2* mutant exuded germination stimulants at slightly higher levels than the wild type (Fig. 4c). Collectively, these results suggest that strigolactones are biosynthesized from carotenoid cleavage products by CCD7 and CD8 and inhibit shoot branching through the MAX-dependent pathway in *Arabidopsis*.

d10 roots are infected by fewer *Striga* plants

Here we have identified strigolactone-deficient and -insensitive mutants. To explore the impact of altered strigolactone levels on the interaction with parasitic weeds, we used rice *d* mutants to observe germination, infection and developmental processes of *S. hermonthica* plants. *S. hermonthica* is an obligate root parasite that infects cereals[34,35], including rice (Fig. 5a). In the vicinity of *d10-1* roots, fewer seeds germinated than did those co-incubated with wild-type or *d3-1* roots (Fig. 5b), consistent with the finding that *d10-1* roots exude lower levels of strigolactones (Fig. 2d, f). As a consequence of the reduced germination frequency, fewer *S. hermonthica* plants established parasitism with *d10-1* mutants in 2 weeks than with wild type or *d3-1* (Fig. 5b). When *S. hermonthica* seeds were co-incubated with *d10-1* seedlings after the induction of germination by (+)-strigol, there was no significant difference in the frequency of successful parasitism among the three genotypes (Fig. 5c). Albeit at a very low frequency, some *S. hermonthica* seeds germinated in the vicinity of *d10-1* roots in the absence of (+)-strigol and then achieved successful infection. Together, these results indicate that compared to wild-type roots, fewer *S. hermonthica* plants can infect *d10-1* roots principally due to lower levels of germination stimulants released from this host. Our results also suggest that, once the *S. hermonthica* seeds germinate, strigolactone deficiency does not significantly affect the subsequent infection processes. We cannot rule out the possibility

that a small amount of strigolactone owing to residual CCD8 activity might exist in the *d10-1* mutant and affect the germination and infection of *S. hermonthica*, because the *d10-1* mutation results in a single amino acid substitution (Supplementary Fig. 1) and may not be a null allele.

Discussion

Outgrowth of axillary buds is, in part, regulated by the interaction of multiple hormonal signals[15]; auxin is actively transported downwards in the shoots and inhibits bud outgrowth, whereas cytokinins move upwards in plants and activate bud outgrowth. In this study, we have shown that the *d* and *max* branching mutants of rice and *Arabidopsis* are deficient in or insensitive to strigolactones, and that exogenously applied strigolactones inhibit shoot branching. Thus, we propose that strigolactones or downstream metabolites act as the long-sought-after hormones in the *D/MAX* pathway. However, it should be noted that the bioactive form(s) of this new hormone class has not been clarified in the current study. Extensive survey of natural strigolactones as seed germination stimulants of root parasites and hyphal branching inducers of arbuscular mycorrhizal fungi revealed highly diverse structures, attributable to modifications on ring ABC and the C2′ configuration[29,36] (Fig. 1b). Moreover, it is not known how these diverse strigolactones are further metabolized in plants. Elucidation of the bioactive form(s) of the branch-inhibiting hormones is a critical next question to explore the distribution, movement and perception of this chemical signal in plants. Considering the chemical structures, more enzymes are likely to be required for

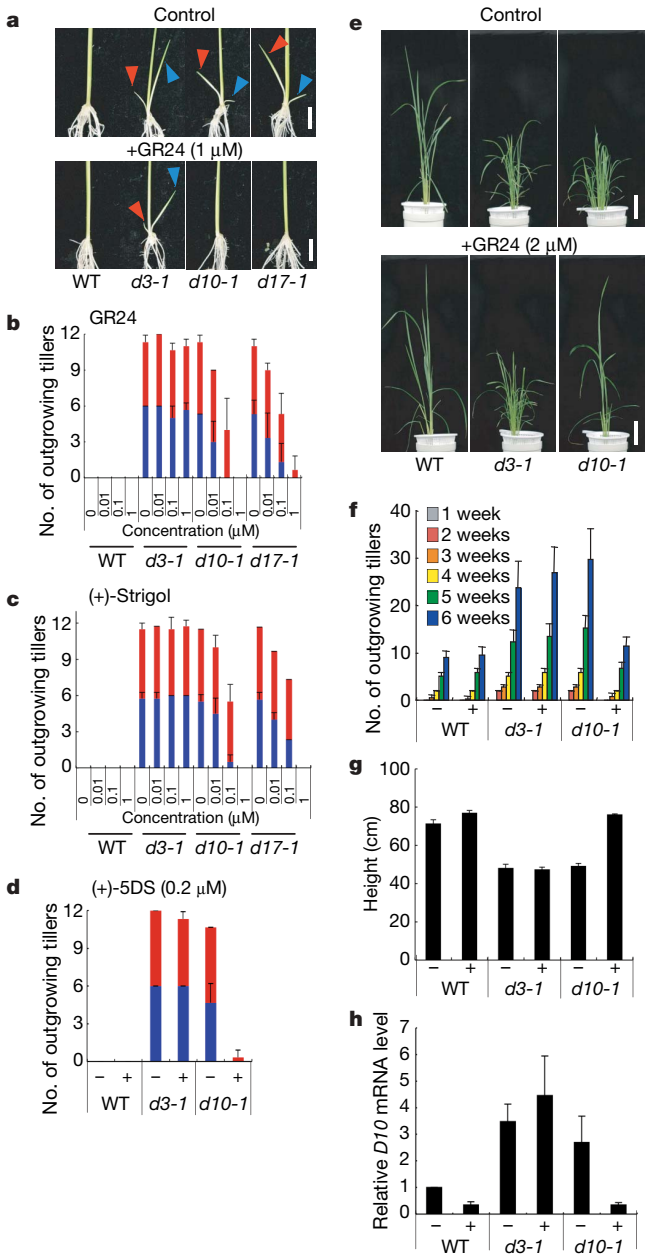


Figure 3 | Effect of strigolactones on rice tillering. **a**–**d**, Two-week-old wild type (WT) and *d* mutants. Scale bar in **a** indicates 1 cm. Blue and red arrowheads (**a**) or bars (**b**–**d**) indicate first and second tillers, respectively. Total number of tillers (over 2 mm) in six seedlings is shown (mean + s.d., $n = 3$). **e**–**g**, Six-week-old plants with (+) or without (–) 2 μM GR24. Scale bar in **e** indicates 10 cm. Weekly changes in the number of tillers (**f**) and the plant height at the sixth week (**g**) is shown (mean + s.d., $n = 4$). **h**, *D10* transcript levels in roots of 8-day-old seedlings with (+) or without (–) 1 μM (+)-GR24 treatment for 24 h (mean and s.d., $n = 3$).

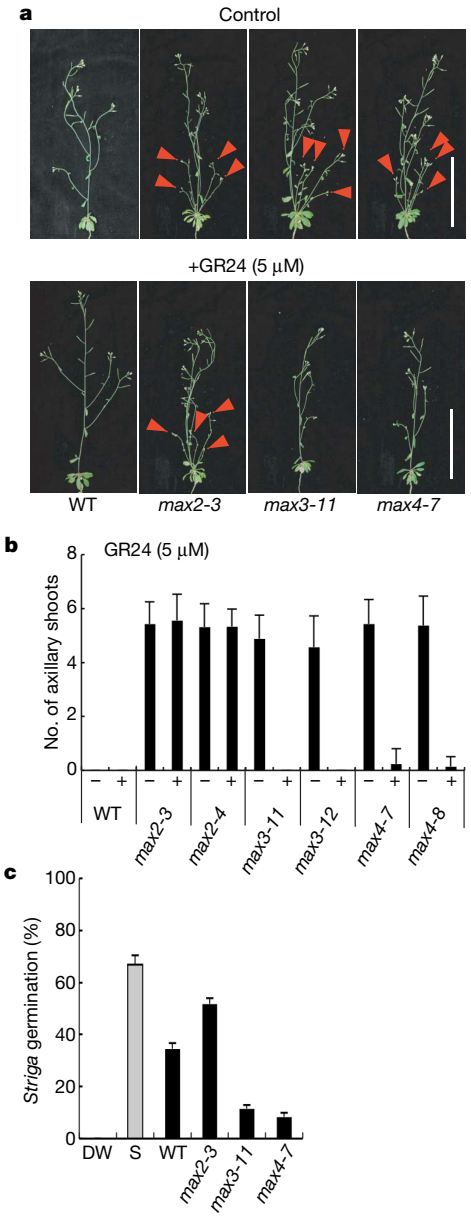


Figure 4 | Effect of GR24 on axillary bud outgrowth of *Arabidopsis*. **a**, Thirty-day-old wild type (WT) and *max* mutants. Red arrowheads indicate outgrowth of axillary buds. Scale bar, 10 cm. **b**, Number of axillary shoots (over 5 mm) is shown (mean + s.d., $n = 12$–16). **c**, Estimation of germination stimulant levels in culture media of 2-week-old seedlings using *Striga* seeds (mean and s.d., $n = 3$). DW, distilled water; S, (+)-strigol (10^{-7} M).

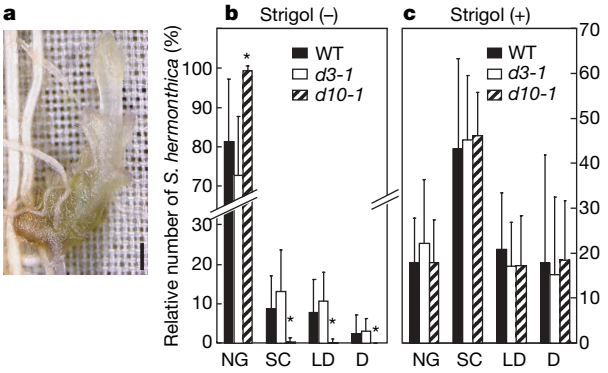


Figure 5 | Infection of rice *d* mutants by *Striga hermonthica*. **a**, *Striga* parasitizing to rice roots 4 weeks after inoculation. Scale bar, 500 μm. **b**, **c**, Ratio of *Striga* plants at each developmental stage 2 weeks after the inoculation of (+)-strigol-treated (**c**) or non-treated (**b**) seeds (mean and s.d.; **b**, $n = 16–17$; **c**, $n = 23–24$). D, died after penetration; LD, leaf developed after the establishment of parasitism; NG, no germination; SC, penetration succeeded and seed coat remained attached; WT, wild type. Asterisks indicate significantly different from wild type (Student's *t*-test, $P < 0.05$).

the biosynthesis of strigolactones, in addition to CCD7, CCD8 and MAX1 (Fig. 1a). Identification of genetic loci defined by additional branching mutants[14,15] may reveal new enzymes in the strigolactone biosynthesis pathway.

Shoot branching is influenced by a wide range of environmental signals[37]. Our findings suggest that strigolactones may have a principal role in mediating the detection of nutrient availability by roots and the resulting alterations in shoot architecture, provided that strigolactone levels are increased in response to inorganic phosphate deficiency, particularly in hosts of arbuscular mycorrhizal fungi[26,28,30,31] (Fig. 2d–f); on inorganic phosphate (and possibly other nutrients) deficiency, a probable adoptive strategy of plants would be to synthesize strigolactones for minimizing shoot branching and maximizing the symbiotic interaction with arbuscular mycorrhizal fungi that facilitate the uptake of mineral nutrients. Seeds of root parasitic plants abuse these chemical signals secreted for the successful symbiosis with arbuscular mycorrhizal fungi to find their potential hosts in soil.

In many parts of the world, the parasitic weeds *Striga* and *Orobanche* are serious agricultural pests[34,35]. Strigolactones have been an important target for parasitic weed control in generating low-germination-stimulant varieties[21]. Although strigolactones have been chemically recognized for decades, the biosynthetic pathway had not been genetically defined. The identification of several *D/MAX* loci as strigolactone biosynthesis genes now allows us to take a first step towards designing new varieties with reduced risk of parasite infections in molecular breeding. In fact, our results show that, at least in an experimental condition, the rice *d10-1* mutant is infected by significantly fewer *S. hermonthica* plants in comparison with wild-type plants, as a consequence of decreased germination frequency of the parasite seeds near the host root (Fig. 5). The use of strigolactone-deficient mutants will also facilitate our understanding of the exact roles of this class of terpenes in communication with arbuscular mycorrhizal fungi in the rhizosphere.

METHODS SUMMARY

Plant materials. Rice and *Arabidopsis* mutants used in this study are shown in Supplementary Fig. 1. Mutations in new mutant alleles used for this study were determined by DNA sequencing. Genotyping was carried out by a PCR-based method using the primers listed in Supplementary Table 1.

Growth conditions and strigolactone treatment. Rice and *Arabidopsis* seeds were surface-sterilized and the seedlings were first grown aseptically on agar media. Plants were then grown hydroponically in growth chambers. For both rice and *Arabidopsis*, strigolactones were added to the hydroponic culture medium.

Strigolactone analysis. The levels of strigolactones released to hydroponic culture media were estimated by germination-stimulating activity using *S. hermonthica* seeds as described previously[32]. Strigolactones were identified and quantified on LC/MS–MS by comparing the retention time and full-scan spectrum with those of authentic standards. We synthesized deuterium-labelled *epi*-5DS (d_1-*epi*-5DS) and used it as an internal standard for quantitative analysis using LC/MS–MS.

Gene expression analysis. We performed qRT–PCR analysis to determine *D10* transcript levels, according to the method described previously[38].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 14 June; accepted 21 July 2008.
Published online 10 August 2008.

1. Beveridge, C. A., Ross, J. J. & Murfet, I. C. Branching mutant *rms-2* in *Pisum sativum*. Grafting studies and endogenous indole-3-acetic acid levels. *Plant Physiol.* **104,** 953–959 (1994).
2. Beveridge, C. A., Ross, J. J. & Murfet, I. C. Branching in pea (Action of Genes *Rms3* and *Rms4*). *Plant Physiol.* **110,** 859–865 (1996).
3. Beveridge, C. A., Symons, G. M. & Turnbull, C. G. Auxin inhibition of decapitation-induced branching is dependent on graft-transmissible signals regulated by genes *Rms1* and *Rms2*. *Plant Physiol.* **123,** 689–697 (2000).
4. Johnson, X. *et al.* Branching genes are conserved across species. Genes controlling a novel signal in pea are coregulated by other long-distance signals. *Plant Physiol.* **142,** 1014–1026 (2006).
5. Stirnberg, P., van De Sande, K. & Leyser, O. *MAX1* and *MAX2* control shoot lateral branching in *Arabidopsis*. *Development* **129,** 1131–1141 (2002).
6. Sorefan, K. *et al. MAX4* and *RMS1* are orthologous dioxygenase-like genes that regulate shoot branching in *Arabidopsis* and pea. *Genes Dev.* **17,** 1469–1474 (2003).
7. Booker, J. *et al.* MAX3/CCD7 is a carotenoid cleavage dioxygenase required for the synthesis of a novel plant signaling molecule. *Curr. Biol.* **14,** 1232–1238 (2004).
8. Booker, J. *et al. MAX1* encodes a cytochrome P450 family member that acts downstream of *MAX3/4* to produce a carotenoid-derived branch-inhibiting hormone. *Dev. Cell* **8,** 443–449 (2005).
9. Turnbull, C. G., Booker, J. P. & Leyser, O. Micrografting techniques for testing long-distance signalling in *Arabidopsis*. *Plant J.* **32,** 255–262 (2002).
10. Snowden, K. C. *et al.* The *Decreased apical dominance1/Petunia hybrida CAROTENOID CLEAVAGE DIOXYGENASE8* gene affects branch production and plays a role in leaf senescence, root growth, and flower development. *Plant Cell* **17,** 746–759 (2005).
11. Simons, J. L., Napoli, C. A., Janssen, B. J., Plummer, K. M. & Snowden, K. C. Analysis of the *DECREASED APICAL DOMINANCE* genes of petunia in the control of axillary branching. *Plant Physiol.* **143,** 697–706 (2007).
12. Ishikawa, S. *et al.* Suppression of tiller bud activity in tillering dwarf mutants of rice. *Plant Cell Physiol.* **46,** 79–86 (2005).
13. Zou, J. *et al.* The rice *HIGH-TILLERING DWARF1* encoding an ortholog of *Arabidopsis MAX3* is required for negative regulation of the outgrowth of axillary buds. *Plant J.* **48,** 687–698 (2006).
14. Arite, T. *et al. DWARF10*, an *RMS1/MAX4/DAD1* ortholog, controls lateral bud outgrowth in rice. *Plant J.* **51,** 1019–1029 (2007).
15. Ongaro, V. & Leyser, O. Hormonal control of shoot branching. *J. Exp. Bot.* **59,** 67–74 (2008).
16. Schwartz, S. H., Qin, X. & Loewen, M. C. The biochemical characterization of two carotenoid cleavage enzymes from *Arabidopsis* indicates that a carotenoid-derived compound inhibits lateral branching. *J. Biol. Chem.* **279,** 46940–46945 (2004).
17. Auldridge, M. E. *et al.* Characterization of three members of the *Arabidopsis* carotenoid cleavage dioxygenase family demonstrates the divergent roles of this multifunctional enzyme family. *Plant J.* **45,** 982–993 (2006).
18. Lechner, E., Achard, P., Vansiri, A., Potuschak, T. & Genschik, P. F-box proteins everywhere. *Curr. Opin. Plant Biol.* **9,** 631–638 (2006).
19. Cook, C. E. *et al.* Germination stimulants II. The structure of strigol–a potent seed germination stimulant for witchweed (*Striga lutea* Lour.). *J. Am. Chem. Soc.* **94,** 6198–6199 (1972).
20. Humphrey, A. J. & Beale, M. H. Strigol: Biogenesis and physiological activity. *Phytochemistry* **67,** 636–640 (2006).
21. Bouwmeester, H. J., Matusova, R., Zhongkui, S. & Beale, M. H. Secondary metabolite signalling in host-parasitic plant interactions. *Curr. Opin. Plant Biol.* **6,** 358–364 (2003).
22. Akiyama, K., Matsuzaki, K. & Hayashi, H. Plant sesquiterpenes induce hyphal branching in arbuscular mycorrhizal fungi. *Nature* **435,** 824–827 (2005).
23. Bradow, J. M., Connick, W. J. Jr, Pepperman, A. B. & Wartelle, L. H. Germination stimulation in wild oats (*Avena fatua* L.) by synthetic strigol analogues and gibberellic acid. *J. Plant Growth Regul.* **9,** 35–41 (1990).
24. Bradow, J. M., Connick, W. J. & Pepperman, A. B. Comparison of the seed germination effects of synthetic analogs of strigol, gibberellic acid, cytokinins and other plant growth regulators. *J. Plant Growth Regul.* **7,** 227–239 (1988).
25. Goldwasser, Y., Yoneyama, K., Xie, X. & Yoneyama, K. Production of strigolactones by *Arabidopsis thaliana* responsible for *Orobanche aegyptiaca* seed germination. *Plant Growth Regul.* **55,** 21–28 (2008).

26. Yoneyama, K. *et al.* Strigolactones, host recognition signals for root parasitic plants and arbuscular mycorrhizal fungi, from Fabaceae plants. *New Phytol.* **179,** 484–494 (2008).
27. Matusova, R. *et al.* The strigolactone germination stimulants of the plant-parasitic *Striga* and *Orobanche* spp. are derived from the carotenoid pathway. *Plant Physiol.* **139,** 920–934 (2005).
28. López-Ráez, J. A. *et al.* Tomato strigolactones are derived from carotenoids and their biosynthesis is promoted by phosphate starvation. *New Phytol.* **178,** 863–874 (2008).
29. Bouwmeester, H. J., Roux, C., Lopez-Raez, J. A. & Bécard, G. Rhizosphere communication of plants, parasitic plants and AM fungi. *Trends Plant Sci.* **12,** 224–230 (2007).
30. Yoneyama, K. *et al.* Nitrogen deficiency as well as phosphorus deficiency in sorghum promotes the production and exudation of 5-deoxystrigol, the host recognition signal for arbuscular mycorrhizal fungi and root parasites. *Planta* **227,** 125–132 (2007).
31. Yoneyama, K., Yoneyama, K., Takeuchi, Y. & Sekimoto, H. Phosphorus deficiency in red clover promotes exudation of orobanchol, the signal for mycorrhizal symbionts and germination stimulant for root parasites. *Planta* **225,** 1031–1038 (2007).
32. Sugimoto, Y. & Ueyama, T. Production of (+)-5-deoxystrigol by *Lotus japonicus* root culture. *Phytochemistry* **69,** 212–217 (2008).
33. Zou, J. *et al.* Characterizations and fine mapping of a mutant gene for high tillering and dwarf in rice (*Oryza sativa* L.). *Planta* **222,** 604–612 (2005).
34. Gressel, J. *et al.* Major heretofore intractable biotic constraints to Africa food security that may be amenable to novel biotechnological solutions. *Crop Prot.* **23,** 661–689 (2004).
35. Joel, D. M. The long-term approach to parasitic weeds control: manipulation of specific developmental mechanisms of the parasite. *Crop Prot.* **19,** 753–758 (2000).
36. Xie, X. *et al.* 2′-Epi-orobanchol and solanacol, two unique strigolactones, germination stimulants for root parasitic weeds, produced by tobacco. *J. Agric. Food Chem.* **55,** 8067–8072 (2007).
37. Cline, M. G. Apical dominance. *Bot. Rev.* **57,** 318–358 (1991).
38. Magome, H., Yamaguchi, S., Hanada, A., Kamiya, Y. & Oda, K. *dwarf* and *delayed-flowering 1*, a novel *Arabidopsis* mutant deficient in gibberellin biosynthesis because of overexpression of a putative AP2 transcription factor. *Plant J.* **37,** 720–729 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to S. Ishikawa for sequencing the *d17-1* allele; K. Fujiwara for assistance in preparing plant materials; N. Makita and H. Sakakibara for their advice on rice hydroponic culture; and Y. Tsuchiya for advice on germination assays. We thank the Salk Institute and the *Arabidopsis* Biological Resource Center for providing *Arabidopsis* T-DNA insertion lines; T. Yokota, K. Yoneyama and X. Xie for sharing information on strigolactone analysis; M. Maekawa for propagating rice seeds; and K. Mori, P. McCourt and A. Gabar Babiker for providing (+)-strigol and 2′-*epi*-orobanchol, (+)-GR24, and *S. hermonthica* seeds, respectively. This work was supported in part by grants from the MEXT of Japan (1820810 to K.Y., 19678001 to K.S. and 19780040 to Sa.Y.) and the MAFF of Japan (Genomics for Agricultural Innovation, IPG0001 to J.K.). M.U. is supported by the RIKEN Special Postdoctoral Researchers Program.

Author Contributions M.U. and T.A. developed and performed rice branching assays. T.A. and J.K. prepared rice genetic materials. M.U. performed *Arabidopsis* branching assays. A.H. carried out LC/MS–MS analysis. Sa.Y. and K.S. designed and performed *S. hermonthica* infection assays. K.A. synthesized labelled *epi*-5DS. N.T.-K. and H.M. carried out qRT–PCR analysis. K.Y. and K.A. provided strigolactones and assisted strigolactone analysis by A.H. Y.K., K.S., K.Y. and J.K. contributed to the experimental design. Sh.Y. directed the project and designed the experiments. Sh.Y. and M.U. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to Sh.Y. (shinjiro@postman.riken.jp).

METHODS

Rice hydroponic culture. We used rice normal cultivars (*Oryza sativa* L. cv. Shiokari and cv. Nipponbare) and tillering dwarf mutants (Supplementary Fig. 1) in this study. Rice seeds were washed in 70% ethanol for 30 s, sterilized in 2.5% sodium hypochlorite solution for 15 min, rinsed with sterile water, and then incubated in sterile water at 28 °C in the dark for 2 days. Germinated seeds were transferred into hydroponic culture media[39] solidified with 0.6% agar (pH 5.7) and cultured at 25 °C under fluorescence white light (150–200 μmol m^{-2} s^{-1}) with a 16 h light/8 h dark photoperiod for 5 days. Each seedling was then transferred to a glass vial containing a sterilized hydroponic culture solution (13 ml), fixed with a piece of sponge at the root–shoot junction to the top of the vial, and grown under the same condition for an additional 7 days (total 2 weeks). The hydroponic solution was supplemented every 3 days. For large-scale cultures, the 2-week-old seedlings were transferred into a 4-l porcelain pot containing the same hydroponic solution and grown under the same condition. After the transfer to pots, the solution was renewed weekly.

***Arabidopsis* hydroponic culture.** We used *Arabidopsis thaliana* ecotype Col-0 as the wild type and *max* mutants (Supplementary Fig. 1) in this study. Seeds were sterilized in 1% sodium hypochlorite solution for 5 min, rinsed with sterile water, and stratified for one day at 4 °C. The seeds were placed on the half-strength Murashige and Skoog (MS) medium[40] containing 1% sucrose and 0.8% agar (pH 5.7) at 22 °C under fluorescence white light (60–70 μmol m^{-2} s^{-1}) with a 16 h light/8 h dark photoperiod for 15 days. Plants were then transferred to a glass pot containing 400 ml hydroponic solution[41] and grown under the same environmental condition for an additional 15 days. The solution was renewed every 3 days. To measure germination stimulants, sterilized and stratified seeds were placed on glass beads (30 ml) wetted with 1/10 strength MS liquid media (10 ml) in a Petri dish (9 cm diameter) and grown for 14 days under the same conditions above. The culture media were collected and subjected to *S. hermonthica* germination assay.

LC/MS–MS analysis. The hydroponic culture media were collected and extracted with ethyl acetate twice after adding d_1-*epi*-5DS as an internal standard. The ethyl acetate phase was concentrated *in vacuo* after drying over sodium sulphate. The roots were homogenized in acetone containing d_1-*epi*-5DS. The filtrates were dried up under nitrogen gas and dissolved in 10% acetone. The extracts were loaded onto Oasis HLB 3 ml cartridges (Waters) and eluted with acetone after washing with de-ionized water. The eluates were loaded onto Sep-pak Silica 1 ml cartridges (Waters), washed with ethyl acetate:*n*-hexane (15:85) and then eluted with ethyl acetate:*n*-hexane (35:65). The *epi*-5DS-containing fractions from culture media and roots were dissolved in 50% acetonitrile and subjected to LC/MS–MS analysis using a system consisting of a quadruple/time-of-flight tandem mass spectrometer (Q-Tof Premier, Waters) and an Acquity Ultra Performance liquid chromatograph (Waters) equipped with a reverse-phase column (Acquity UPLC BEH-C_{18}, 2.1 × 50 mm, 1.7 μm; Waters). The mobile phase was changed from 30% acetonitrile containing 0.05% acetic acid to 40% and 70% in 5 and 10 min after the injection, respectively, at a flow rate of 0.2 ml min^{-1}. Data analysis was performed as we described previously for gibberellin analysis using MassLynx software (v. 4.1)[42].

Chemicals. GR24, 5DS and 5DS isomers were synthesized as described previously[22,43]. (+)-Strigol and 2′-*epi*-orobanchol were provided by K. Mori. For experiments in Fig. 3h, we used (+)-GR24 (courtesy of P. McCourt). The synthesis of d_1-(*epi*)-5DS was carried out as described previously for non-labelled 5DS[22]. The ABC ring was formylated with deuterium-labelled methyl formate and the following alkylation with racemic 4-bromo-2-methyl-2-buten-4-olide provided [6′-d]-5DS and its 2′-epimer (Supplementary Fig. 2). (±)-[6′-d_1]-*epi*-5DS was purified by a silica gel column (Wakogel C-200, Wako Pure Industries; *n*-hexane-ethyl acetate stepwise) and semi-preparative high-performance liquid chromatography on reverse-phase (Inertsil ODS-3, GL Sciences; 70% acetonitrile in water) and normal-phase (Inertsil SIL-100A, GL Sciences; 15% ethanol in *n*-hexane) columns.

Germination assay. Germination assays using *S. hermonthica* were performed as described previously[32]. For each bioassay, de-ionized water and (+)-strigol solution were used as negative and positive controls, respectively.

Gene expression analysis. Total RNA was extracted from roots using the RNeasy Maxi kit (Qiagen). qRT–PCR was carried out to determine *D10* transcript levels using gene-specific primers and a Taq-Man probe (Supplementary Table 1). Ubiquitin expression was used as an internal standard.

***S. hermonthica* infection assay.** *S. hermonthica* infections were analysed using a rhizotron system as described previously[44], with slight modifications. Briefly, 1-week-old rice seedlings were transferred to root-observing rhizotron chambers (225 mm × 225 mm Petri dish filled with rockwool and nylon mesh) supplied with 50 ml half-strength MS media, and grown for 2 weeks in a green house with a 12 h photoperiod (170–450 μmol m^{-2} s^{-1}) at day/night temperature cycles of 28 °C/20 °C. *S. hermonthica* seeds were preconditioned on moist glass fibre filter papers (GF/A, Wattman) at 26 °C in the dark for 2 weeks, and treated with or without 10^{-9} M (+)-strigol for 5 h in the dark. After rinsing with excess water, approximately 50 parasite seeds were carefully placed along rice roots and the rhizotrons were incubated under the same growth conditions described above. The status of germination, infection and development of *S. hermonthica* was evaluated after 2 and 4 weeks of co-cultivation.

39. Kamachi, K., Yamaya, T., Mae, T. & Ojima, K. A role for glutamine synthetase in the recombination of leaf nitrogen during natural senescence in rice leaves. *Plant Physiol.* **96,** 411–417 (1991).
40. Murashige, T. & Skoog, F. A revised medium for rapid growth and bioassays with tobacco tissue cultures. *Physiol. Plant.* **15,** 473–497 (1962).
41. Norén, H., Svensson, P. & Andersson, B. A convenient and versatile hydroponic cultivation system for *Arabidopsis thaliana*. *Physiol. Plant.* **121,** 343–348 (2004).
42. Varbanova, M. *et al.* Methylation of gibberellins by *Arabidopsis* GAMT1 and GAMT2. *Plant Cell* **19,** 32–45 (2007).
43. Mangnus, E. M., Jan Dommerholt, F., de Jong, R. L. P. & Zwaneburg, B. Improved synthesis of strigol analogue GR24 and evaluation of the biological activity of its diastereomers. *J. Agric. Food Chem.* **40,** 1230–1235 (1992).
44. Gurney, A. L., Slate, J., Press, C. & Scholes, J. D. A novel form of resistance in rice to the angiosperm parasite *Striga hermonthica*. *New Phytol.* **169,** 199–208 (2006).

LETTERS

A blast wave from the 1843 eruption of η Carinae

Nathan Smith[1]

Very massive stars shed much of their mass in violent precursor eruptions[1] as luminous blue variables[2] (LBVs) before reaching their most likely end as supernovae, but the cause of LBV eruptions is unknown. The nineteenth-century eruption of η Carinae, the prototype of these events[3], ejected about 12 solar masses at speeds of 650 km s^{-1}, with a kinetic energy of almost 10^{50} erg (ref. 4). Some faster material with speeds up to 1,000–2,000 km s^{-1} had previously been reported[5–8] but its full distribution was unknown. Here I report observations of much faster material with speeds up to 3,500–6,000 km s^{-1}, reaching farther from the star than the fastest material in previous reports[5]. This fast material roughly doubles the kinetic energy of the nineteenth-century event and suggests that it released a blast wave now propagating ahead of the massive ejecta. As a result, η Carinae's outer shell now mimics a low-energy supernova remnant. The eruption has usually been discussed in terms of an extreme wind driven by the star's luminosity[2,3,9,10], but the fast material reported here indicates that it may have been powered by a deep-seated explosion rivalling a supernova, perhaps triggered by the pulsational pair instability[11]. This may alter interpretations of similar events seen in other galaxies.

η Carinae[3] is the most luminous and the most extensively studied of the LBVs[1,2]. It is the most massive and luminous star in our region of the Milky Way, and it provides important constraints on the pre-supernova evolution of the most massive stars, even if it is a rather extreme example of the instabilities that they encounter. η Carinae probably began its life with an initial mass of about 150 solar masses ($M_{\odot}$), and has a current estimated mass of about 90–100$M_{\odot}$, with much of the difference lost in sudden giant eruptions in the past few thousand years[1]. It is orbited by a companion star[12], but it is unclear whether the companion had a role in its eruptive instability[4].

It is now well established that the so-called 'Homunculus' nebula was ejected during the 1840s near the peak of η Carinae's eruption[13]. That event ejected at least 12$M_{\odot}$ of gas moving at speeds of up to 650 km s^{-1} (refs 4, 14). This speed roughly matches the present-day polar wind speed[15] and is close to the expected escape velocity from the star's surface. The observed ejecta speeds reported here are much faster than this, and are quite surprising because such high speeds are not expected in a wind from an evolved massive star such as η Carinae.

The data reported here come in two varieties. First, near-infrared spectra of the He I λ10830 emission line obtained with the GNIRS instrument on the Gemini South telescope (Fig. 1) reveal emission from what seems to be a fast outflowing disk of material near the waist of bipolar lobes outside the Homunculus, with a spatial scale and expansion speeds roughly twice those of the well-known equatorial skirt of the Homunculus. Although extended He I λ10830 emission had been discovered at one position outside the Homunculus in a previous study[6], the larger structure and faster velocities reported here have not previously been seen in any data. This material is flowing away from the star with radial speeds of roughly 1,000–2,000 km s^{-1} and seems to be present all the way around the Homunculus not far from the equatorial plane.

Second, visual-wavelength spectra of the region near Hα show [N II] λ6548 and λ6583 emission from extremely fast material with Doppler shifts of roughly −3,000 to +2,500 km s^{-1} (see Fig. 2 and Supplementary Fig. 1). Emission from this fast material was suspected to exist in previous data of lower quality obtained at one position around η Carinae[5], and it provided the main motivation for obtaining the high-quality data seen here. The fast N-rich gas reported here is spatially coincident with or inside the soft X-ray shell around η Carinae (Supplementary Fig. 1). It is apparently running into the many N-rich outer knots seen in visual-wavelength images[5,16–18], because it is near or interior to these knots, but is expanding outwards at a much faster pace. The fastest material has not been seen in most [N II] images of η Carinae, such as those obtained with the Hubble Space Telescope (Fig. 1), because it is Doppler-shifted far out of the narrow-filter bandpass range. It follows the same bipolar expansion pattern seen for the Homunculus[14] as well as that of the slower [N II]-bright knots in the outer ejecta[19] (blueshifted to the southeast, redshifted to the northwest). The observed Doppler shifts of up to 3,000 km s^{-1} are a lower limit to the deprojected maximum velocity of this material. Because this material is seen far off to the side of η Carinae, its trajectory must be inclined away from our line of sight, with plausible values for the true space velocity ranging from 3,500 to 6,000 km s^{-1} (Supplementary Fig. 2). The higher speeds would require that this material has been accelerated by the pressure behind the blast wave since its ejection.

The fastest material has [N II]/Hα ratios similar to the slower N-rich knots[20], indicating that it was launched at these speeds by the same evolved primary star that ejected the N-rich Homunculus[21] and all its other N-rich material. Although it is intriguing that speeds about 3,000 km s^{-1} are similar to the wind speeds of the secondary star in models for the X-ray emission from η Carinae[22], the notion that perhaps the secondary star was responsible for ejecting this fast N-rich material has little merit on closer examination. If one ascribes the 1843 outburst to the as yet unseen secondary star in the η Carinae binary system, several inconsistencies arise. First, the expansion speed, bipolar shape and polar axis orientation of the Homunculus[14] (known to have been ejected in the 1843 event[13]) match the present-day properties of the primary star's wind[15]. Second, the star we observe now as the primary is wildly variable, has an enormous mass-loss rate and shows visible signs of recovery from the 1843 eruption, whereas the putative secondary star has not been detected. Third, for an evolved massive star to produce the wind speeds required of the secondary star, it must be a compact H-poor Wolf–Rayet star, but this is incompatible with η Carinae's multiple ejections of massive H-rich nebulosity. In short, passing the burden of mass ejection to η Carinae's companion star is not viable.

In any case, speeds of this order evoke properties more like supernova explosions than super-Eddington winds[9], and they provide a fundamental new clue to the nature of the trigger behind η Carinae's eruption. Combining the fast near-equatorial material traced by He I λ10830 with fast polar ejecta traced by [N II] implies a bipolar

[1]Astronomy Department, University of California, 601 Campbell Hall, Berkeley, California 94720-3411, USA.

forward shock geometry, perhaps like that depicted in Fig. 3. The geometry is similar to that of the Homunculus itself[7,14], but threefold to fourfold its size and speed. This fast blast wave from the nineteenth-century eruption of η Carinae is powered by the very fast and low-density N-rich material that is coincident with or interior to the X-ray shell[23]. It has been found[23] that the properties of the soft X-ray shell were consistent with a blast-wave interpretation. It is the interaction between the blast wave from the 1843 event overtaking the 500–1,000-year-old[17] clumpy N-rich ejecta shell—and not the older condensations running into ambient material—that apparently gives rise to the X-ray emission and shock ionization of the outer ejecta. This fast blast wave is likely to be important for the total energy budget of the 1843 eruption. If the high-speed material contains only 5% of the mass of the Homunculus, for example (corresponding to a likely average density of at least 500 cm^{-3} filling the volume of a sphere in the range of radii from the star over which the fast material is seen with a ~0.5 filling factor), but is moving at speeds fivefold faster, it may contain 7×10^{49} erg. This amount of kinetic energy would be comparable to, or could even exceed, the kinetic energy of the main Homunculus nebula[4].

What could be the origin of this extremely fast and energetic forward shock? Currently favoured models for the mass loss during a giant LBV eruption involve a radiation-driven wind as the star exceeds the Eddington luminosity limit[2,3,9,10], but in that case one expects a wind to be quite slow[9]. Fast material may escape through regions between dense clumps if the wind becomes highly porous[9], but this still does not explain how material is accelerated to speeds far exceeding the star's escape speed. Instead, an event of this sort, which also launched the $12M_{\odot}$ of matter in the Homunculus[4], may have been a deep-seated explosion. Its constraints are reminiscent of the pulsational pair instability[11] or some other instability associated with explosive nuclear burning in the latest stages of evolution. However, those burning events are expected to occur shortly (~10 years) before the final core collapse supernova explosion, as proposed for the progenitor of the extremely luminous supernova 2006gy[24,25], or in some cases as much as 1,000 years before core collapse. In the unique case of supernova 2006jc, a precursor LBV-like outburst was actually observed just 2 years before the final supernova[26,27]. η Carinae has not yet reached core collapse even though the eruption happened more than 160 years ago, and it is thought to have undergone similar eruptive events ~1,000 years ago[17].

The presence of a fast blast wave has critical implications for understanding the nature of η Carinae's 1843 eruption, and the

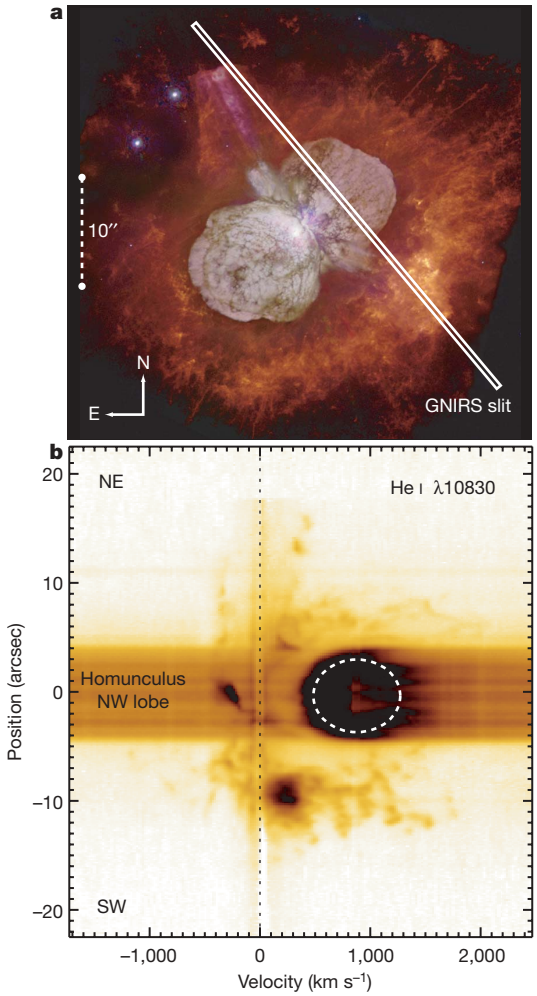


Figure 1 | Example of the velocity structures seen in the He I λ10830 line. **a**, The 25 × 0.3 arcsec^2 spectroscopic slit aperture of the GNIRS spectrograph was oriented at position angle = +40° (SW to NE) crossing the NW polar lobe of the Homunculus nebula. **b**, In the resulting position–velocity plot, the broad horizontal strip is from reflected continuum light scattered by dust in the Homunculus NW polar lobe, and He I λ10830 emission from fast material is seen outside that structure over a large range of velocities. The image in **a** is a composite-colour visual-wavelength image made from several different exposures obtained with the WFPC2 camera on the Hubble Space Telescope, through the F336W (blue), F631N (green) and F658N (red) filters, and is shown here only for a comparison of the ejecta morphology with the slit placement. These data were reduced in a manner consistent with previous similar data[6,14]. The outer gas that emits He I in the spectrum in **b** appears red/orange in this image because of strong [N II] emission, and the position–velocity plot is a subsection of a 1.0–1.15-μm spectrum centred on the bright He I λ10830 emission line, obtained on 2005 March 20 with the GNIRS spectrograph mounted on the Gemini South telescope. It is a combination of four individual exposures of 90 s each. The spatial scale is the same in **a** and **b**. The dashed white ellipse in **b** marks a He I feature that arises from emission formed in the central star's wind but is scattered and red-shifted by dust grains; it is unrelated to the present topic.

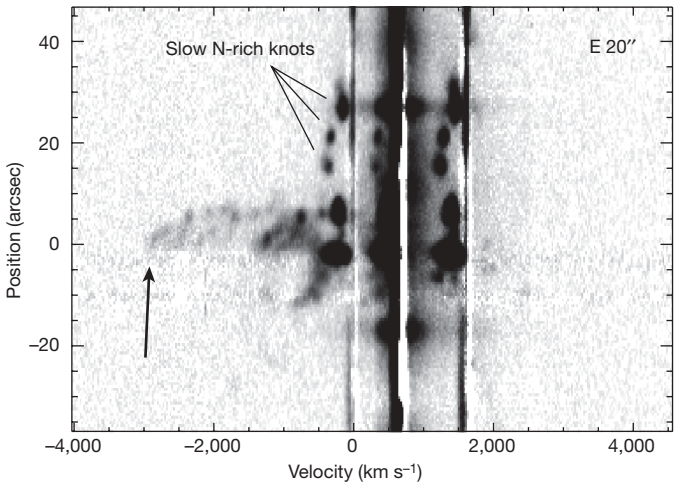


Figure 2 | Extremely fast nitrogen-rich material surrounding η Carinae. This panel shows a two-dimensional position–velocity spectrum at an example offset position 20 arcsec east of the star, resulting from a 15-min exposure obtained on 2006 March 15 with the RC-Spec spectrograph mounted on the Cerro Tololo Inter-American Observatory 4-m telescope. Similar visual-wavelength spectra at several additional offset positions are included in Supplementary Fig. 1. The velocity scale is set for the [N II] λ6548 line to aid the interpretation, because blue-shifted emission dominates at this position. [N II] λ6548 emission reveals material travelling at blue-shifted Doppler speeds of up to 3,000 km s^{-1} (arrowed), coincident with the outer shell seen in X-rays, whereas the dense N-rich knots seen in Hubble Space Telescope images have much lower speeds of 100–300 km s^{-1}. The slow-moving knots have been studied in detail[5,16–19], but the faster material at −2,000 to −3,000 km s^{-1} had only been suspected at one location in a single exposure in previous data of lower quality[5]. These earlier observations provided the motivation for the higher-quality spectra presented here.

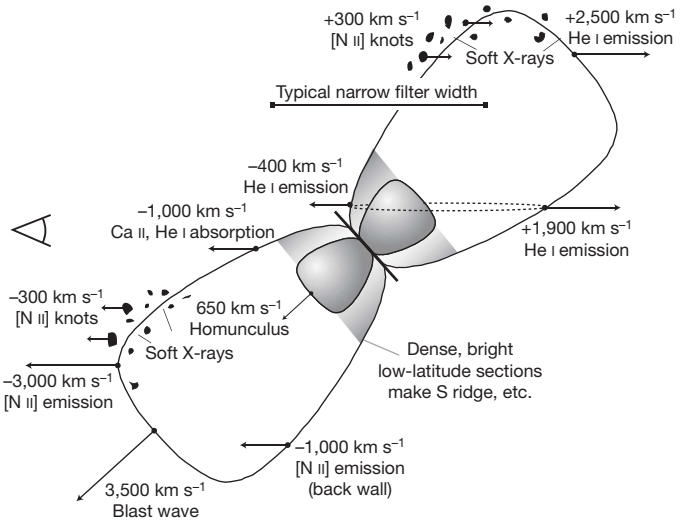


Figure 3 | Illustration of the possible geometry of η Carinae's blast wave. This simplified diagram depicts the basic geometry of η Carinae's outer ejecta, with an Earth-based observer located to the left. The inner shaded structure shows the shape and orientation of the Homunculus nebula[7,14]. Some of the relatively slow, dense N-rich knots from a previous eruptive event ~1,000 years ago[17] are now being overrun by the fast forward shock from the 1843 eruption of η Carinae. This collision gives rise to the soft X-ray shell[23], and ultraviolet radiation from shocked gas ionizes pre-shock material in the slower N-rich knots seen in images. For background information on previous studies of these slower outer debris, see refs 5, 16–20. Part of the thin sidewall of the approaching side of this bipolar shock structure crosses in front of our line of sight to the Homunculus nebula, which is a reflection nebula. This part of the structure, blue-shifted at roughly $-1{,}000\ \mathrm{km\ s^{-1}}$, may previously have been seen in absorption in the Ca II HK (ref. 7) and He I λ10830 (ref. 6) lines, and in emission in Hα and [N II] lines[7,8], but the connection between these velocity features and the polar blast wave shown here is not entirely clear with available data. A typical narrow-band imaging filter, such as the F658N filter of the WFPC2 camera on the Hubble Space Telescope, will exclude the fastest Doppler-shifted features.

LBV instability in general. In effect, it requires that these outbursts can have an explosive aspect that has not previously been appreciated, blurring the observational and phenomenological distinctions between giant LBV eruptions and supernova explosions in the most massive stars. A corollary is that the extended nebula now seen around η Carinae is analogous to a low-energy supernova remnant, and the X-ray emission arising from the interaction between the blast wave and surrounding material may have been stronger in the past. η Carinae is also the prototype for a class of stellar eruptions seen in other galaxies, sometimes called 'supernova impostors', type V supernovae, or faint type IIn supernovae[28]. Indeed, there may even be an ill-defined continuum in energy between eruptions such as η Carinae, super-outbursts of LBVs such as supernova 1961V (ref. 29) and faint core-collapse supernovae. Because these LBV outbursts can evidently have powerful blast waves of a few thousand $\mathrm{km\ s^{-1}}$, the observations reported here indicate that detections of radio synchrotron or X-ray emission are not reliable ways to distinguish between LBV eruptions and genuine core-collapse supernova explosions.

Received 1 April; accepted 17 July 2008.

1. Smith, N. & Owocki, S. P. On the role of continuum-driven eruptions in the evolution of very massive stars and Population III stars. *Astrophys. J.* **645,** L45–L48 (2006).
2. Humphreys, R. M. & Davidson, K. The luminous blue variables: Astrophysical geysers. *Publ. Astron. Soc. Pacif.* **106,** 1025–1051 (1994).
3. Davidson, K. & Humphreys, R. M. Eta Carinae and its environment. *Annu. Rev. Astron. Astrophys.* **35,** 1–32 (1997).
4. Smith, N. *et al.* Mass and kinetic energy of the Homunculus nebula around η Carinae. *Astron. J.* **125,** 1458–1466 (2003).
5. Smith, N. & Morse, J. A. Nitrogen and oxygen abundance variations in the outer ejecta of η Carinae: Evidence for recent chemical enrichment. *Astrophys. J.* **605,** 854–863 (2004).
6. Smith, N. Dissecting the Homunculus nebula around Eta Carinae with spatially resolved near-infrared spectroscopy. *Mon. Not. R. Astron. Soc.* **337,** 1252–1268 (2002).
7. Davidson, K., Smith, N., Gull, T. R., Ishibashi, K. & Hillier, D. J. The shape and orientation of the Homunculus nebula based on spectroscopic velocities. *Astron. J.* **121,** 1569–1577 (2001).
8. Currie, D. G., Dorland, B. N. & Kaufer, A. Discovery of a high velocity, spatially extended emission 'shell' in front of the southeast lobe of the η Carinae Homunculus. *Astron. Astrophys.* **389,** L65–L68 (2002).
9. Owocki, S. P., Gayley, K. G. & Shaviv, N. J. A porosity-length formalism for photon-tiring-limited mass loss from stars above the Eddington limit. *Astrophys. J.* **616,** 525–541 (2004).
10. Shaviv, N. J. The porous atmosphere of η Carinae. *Astrophys. J.* **532,** L137–L140 (2000).
11. Heger, A. & Woosley, S. F. The nucleosynthetic signature of Population III. *Astrophys. J.* **567,** 532–543 (2002).
12. Damineli, A., Conti, P. S. & Lopes, D. F. Eta Carinae: A long-period binary? *N. Astron.* **2,** 107–117 (1997).
13. Morse, J. A. *et al.* Hubble Space Telescope proper-motion measurements of the η Carinae nebula. *Astrophys. J.* **548,** L207–L211 (2001).
14. Smith, N. The structure of the Homunculus. I. Shape and latitude dependence from H_2 and [Fe II] velocity maps of η Carinae. *Astrophys. J.* **644,** 1151–1163 (2006).
15. Smith, N., Davidson, K., Gull, T. R., Ishibashi, K. & Hillier, D. J. Latitude-dependent effects in the stellar wind of η Carinae. *Astrophys. J.* **586,** 432–450 (2003).
16. Walborn, N. R. The complex outer shell of Eta Carinae. *Astrophys. J.* **204,** L17–L19 (1976).
17. Walborn, N. R., Blanco, B. M. & Thackeray, A. D. Proper motions in the outer shell of Eta Carinae. *Astrophys. J.* **219,** 498–503 (1978).
18. Meaburn, J., Wolstencroft, R. D. & Walsh, J. R. Echelle and spectropolarimetric observations of the Eta Carinae nebulosity. *Astron. Astrophys.* **181,** 333–342 (1987).
19. Weis, K., Duschl, W. J. & Bomans, D. J. High velocity structures in, and the X-ray emission from the LBV nebula around η Carinae. *Astron. Astrophys.* **367,** 566–576 (2001).
20. Davidson, K., Dufour, R. J., Walborn, N. R. & Gull, T. R. Ultraviolet and visual wavelength spectroscopy of gas around Eta Carinae. *Astrophys. J.* **305,** 867–879 (1986).
21. Smith, N., Brooks, K. J., Koribalski, B. & Bally, J. Cleaning up η Carinae: Detection of ammonia in the Homunculus nebula. *Astrophys. J.* **645,** L41–L44 (2006).
22. Pittard, J. M. & Corcoran, M. F. In hot pursuit of the hidden companion of η Carinae: An X-ray determination of the wind parameters. *Astron. Astrophys.* **383,** 636–647 (2002).
23. Seward, F. D. *et al.* Early Chandra X-ray observations of η Carinae. *Astrophys. J.* **553,** 832–836 (2001).
24. Woosley, S. F., Blinnikov, S. & Heger, A. Pulsational pair instability as an explanation for the most luminous supernovae. *Nature* **450,** 390–392 (2007).
25. Smith, N. *et al.* SN 2006gy: Discovery of the most luminous supernova ever recorded, powered by the death of an extremely massive star like η Carinae. *Astrophys. J.* **666,** 1116–1128 (2007).
26. Foley, R. J. *et al.* 2006jc: A Wolf–Rayet star exploding in a dense He-rich circumstellar medium. *Astrophys. J.* **657,** L105–L108 (2007).
27. Pastorello, A. *et al.* A giant outburst two years before the core-collapse of a massive star. *Nature* **447,** 829–832 (2007).
28. Van Dyk, S. D. *et al.* SN 1997bs in M66: Another extragalactic η Carinae analog? *Publ. Astron. Soc. Pacif.* **112,** 1532–1541 (2000).
29. Goodrich, R. W., Stringfellow, G. S., Penrod, G. D. & Filippenko, A. V. S. N. 1961V: An extragalactic Eta Carinae analog? *Astrophys. J.* **342,** 908–916 (1989).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements I acknowledge continuing collaboration and relevant discussions with the supernova group at the University of California at Berkeley.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to N.S. (nathans@astro.berkeley.edu).

Vol 455|11 September 2008|doi:10.1038/nature07244

LETTERS

A Mott insulator of fermionic atoms in an optical lattice

Robert Jördens[1]*, Niels Strohmaier[1]*, Kenneth Günter[1,2], Henning Moritz[1] & Tilman Esslinger[1]

Strong interactions between electrons in a solid material can lead to surprising properties. A prime example is the Mott insulator, in which suppression of conductivity occurs as a result of interactions rather than a filled Bloch band[1]. Proximity to the Mott insulating phase in fermionic systems is the origin of many intriguing phenomena in condensed matter physics[2], most notably high-temperature superconductivity[3]. The Hubbard model[4], which encompasses the essential physics of the Mott insulator, also applies to quantum gases trapped in an optical lattice[5,6]. It is therefore now possible to access this regime with tools developed in atomic physics. However, an atomic Mott insulator has so far been realized only with a gas of bosons[7], which lack the rich and peculiar nature of fermions. Here we report the formation of a Mott insulator of a repulsively interacting two-component Fermi gas in an optical lattice. It is identified by three features: a drastic suppression of doubly occupied lattice sites, a strong reduction of the compressibility inferred from the response of double occupancy to an increase in atom number, and the appearance of a gapped mode in the excitation spectrum. Direct control of the interaction strength allows us to compare the Mott insulating regime and the non-interacting regime without changing tunnel-coupling or confinement. Our results pave the way for further studies of the Mott insulator, including spin-ordering and ultimately the question of *d*-wave superfluidity[6,8].

The physics of a Mott insulator is well captured by the celebrated Hubbard model, which is widely used to describe strongly interacting electrons in a solid. It assumes a single static energy band for the electrons and local interactions; that is, spin-up and spin-down fermions are moving on a lattice and interact when occupying the same lattice site. The consequence of strong repulsive interactions is that even fermions in different spin states tend to avoid each other. In a half-filled band the particles get localized, and an incompressible state with one fermion per site forms. Because no symmetry is broken, the transition between the metallic and the Mott insulating regime at finite temperature shows a crossover rather than a phase transition.

The Hubbard model ignores various complexities of materials[2], but it has been highly successful in studying the nature of the Mott insulating regime, including magnetic phenomena[2] and high-temperature superconductivity[3]. However, despite its simplicity, it turned out that the fermionic Hubbard model is in many cases computationally intractable and that important puzzles remain to be solved. In particular, the question of whether the ground state of the lightly doped two-dimensional Hubbard model supports *d*-wave superconductivity is as yet unanswered.

In comparison with real materials, a fermionic quantum gas trapped in an optical lattice is a much purer realization of the Hubbard model[5,6,9–11]. It offers a new approach to understanding the physics of strongly correlated systems. In an optical lattice three mutually perpendicular standing laser waves create a periodic potential for the atoms. The kinetics of the atoms is determined by their tunnelling rate between neighbouring lattice sites, and the interaction is due to interatomic collisions occurring when two atoms are on the same site. In a gas of fermions in different spin states this collisional interaction can be widely tuned through a Feshbach resonance without encountering significant atom losses[10].

A landmark result has been the observation of the transition from superfluid to Mott insulator by using bosonic atoms trapped in an optical lattice[7]. Yet it is the fermionic character combined with repulsive interactions that provides the intimate link to fundamental questions in strongly correlated electron systems. Whereas experimental studies of fermionic quantum gases in three-dimensional optical lattices have so far been scarce and focused on non-interacting and attractively interacting cases[12–16], we investigate the repulsive Fermi–Hubbard model and its distinctive feature, the Mott insulator.

In optical lattice experiments the presence of an underlying harmonic trapping potential has an important influence on the observable physics. Let us first consider a zero-temperature Fermi gas prepared in an equal mixture of two non-interacting spin components. All available single-particle quantum states will be filled up to the Fermi energy and, for a sufficiently large number of trapped atoms, a band insulating region with two atoms per site appears in the trap centre, surrounded by a metallic shell with decreasing filling (Fig. 1).

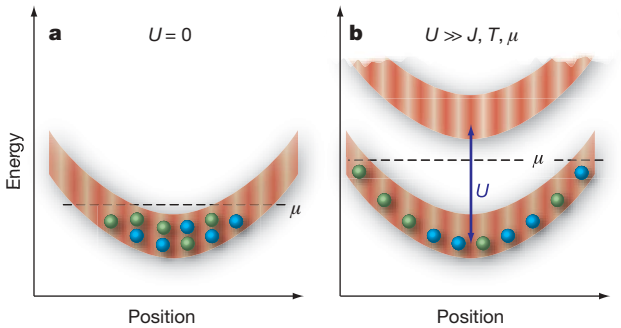


Figure 1 | Energy spectrum of a Fermi gas in an optical lattice with an underlying confining potential. a, In the non-interacting case the curvature of the lowest Bloch band reflects the harmonic confinement. At zero temperature all states up to the chemical potential μ are filled with atoms of both spin states (green and blue). **b**, In the Mott insulating limit the energy cost for creating doubly occupied sites greatly exceeds the temperature T and the kinetic energy parameterized by J, giving rise to a gap of order U. The energy spectrum of single-particle excitations is then depicted by two Hubbard bands. Doubly occupied sites correspond to atoms in the upper Hubbard band.

[1]Institute for Quantum Electronics, ETH Zurich, 8093 Zurich, Switzerland. [2]Laboratoire Kastler Brossel, École Normale Supérieure, 24 rue Lhomond, 75005 Paris, France.
*These authors contributed equally to this work.

An important quantity with which to characterize the state of the system is the fraction D of atoms residing on lattice sites that are occupied by two atoms, one from each component. For the non-interacting case this double occupancy should increase in a continuous fashion with the number N of atoms in the trap.

A very different behaviour can be predicted for a gas with increasingly strong repulsive interactions. A Mott insulator will appear[17,18], at first in those regions of the trap where the local filling is approximately one atom per site. For very strong repulsion the entire centre of the trap contains a Mott insulating phase and double occupancy is suppressed (Fig. 1). Because the Mott insulating region is incompressible[18,19], the suppression of double occupancy should be robust against a tightening of the trapping potential or, equivalently, against an increase in the number of trapped atoms. However, once the chemical potential μ has reached a level at which double occupation of sites becomes favourable, a metallic phase appears in the centre and the double occupancy increases accordingly. The energy spectrum in the Mott insulating phase is gapped, with a finite energy cost required to bring two atoms onto the same lattice site. This energy has to be large in comparison with the temperature, to keep the number of thermally excited doubly occupied sites small. Thermally excited holes in the centre are suppressed by the chemical potential μ.

Our experiment is performed with a quantum degenerate gas of fermionic ^{40}K atoms, prepared in a balanced mixture of two magnetic sublevels of the $F = 9/2$ hyperfine manifold, where F is the total angular momentum. Feshbach resonances allow us to tune the s-wave scattering length between $a = (240 \pm 4)a_0$ and $(810 \pm 40)a_0$ as well as to prepare non-interacting samples. Here a_0 is the Bohr radius. The two-component Fermi gas is subjected to the potential of a three-dimensional optical lattice of simple cubic symmetry. In terms of the recoil energy $E_r = h^2/(2m\lambda^2)$, the lattice potential depth V_0 is chosen between 6.5 E_r and 12E_r. Here h is Planck's constant, m is the atomic mass and $\lambda = 1{,}064$ nm is the wavelength of the lattice beams. The system is described by the Hubbard Hamiltonian

$$\hat{H} = -J \sum_{\langle ij \rangle, \sigma} \left(\hat{c}^{\dagger}_{j\sigma} \hat{c}_{i\sigma} + \text{h.c.} \right) + U \sum_i \hat{n}_{i\uparrow} \hat{n}_{i\downarrow} + \sum_i \varepsilon_i \hat{n}_i .$$

The onsite interaction energy is given by U and the tunnelling matrix element between nearest neighbours $\langle ij \rangle$ by J. The quotient $U/(6J)$ that characterizes the ratio between interaction and kinetic energy can be tuned from zero to a maximum value of 30. The fermionic creation operator for an atom on the lattice site i is given by $\hat{c}^{\dagger}_{i\sigma}$, where $\sigma \in \{\uparrow, \downarrow\}$ denotes the magnetic sublevel and h.c. is the Hermitian conjugate. The particle number operator is $\hat{n}_i = \hat{n}_{i\uparrow} + \hat{n}_{i\downarrow}$, $\hat{n}_{i\sigma} = \hat{c}^{\dagger}_{i\sigma} \hat{c}_{i\sigma}$, and ε_i is the energy offset experienced by an atom on lattice site i due to the harmonic confining potential.

To characterize the state of the Fermi gas in the optical lattice we have developed a technique to reliably measure the fraction D of atoms residing on doubly occupied sites down to values of 1%. The experimental procedure is as follows. The depth of the optical lattice is rapidly increased to 30E_r to inhibit further tunnelling. In the next step we shift the energy of atoms on doubly occupied sites by approaching a Feshbach resonance. This enables us to specifically address only atoms on doubly occupied sites by using a radio-frequency pulse to transfer one of the spin components to a third, previously unpopulated magnetic sublevel. The fraction of transferred atoms is obtained from absorption images and allows us to deduce the double occupancy.

The double occupancy as a function of total atom number is plotted in Fig. 2a, where the non-interacting situation is compared with the case of strong repulsive interactions. The former shows the expected rapid increase of double occupancy with atom number. A strikingly different behaviour is observed in the strongly repulsive regime with $U \gg J$, T, μ, where a Mott insulator is expected. The double occupancy is strongly reduced to values systematically below

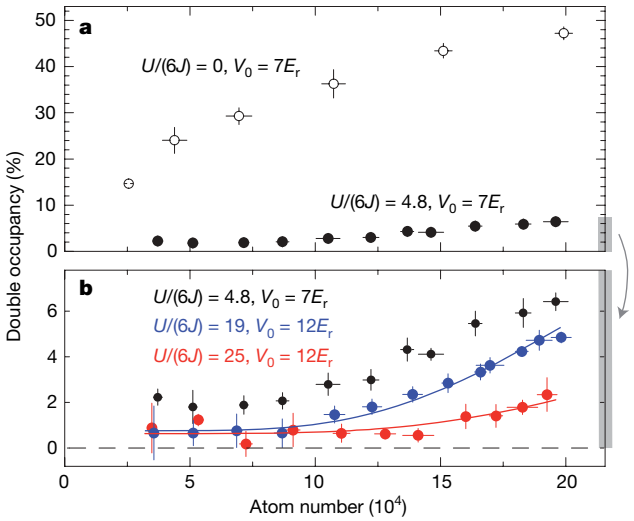


Figure 2 | Double occupancy in the non-interacting and Mott insulating regimes. a, A significant increase in the double occupancy with atom number is observed in the non-interacting regime (open circles), whereas on entering the Mott insulating regime the double occupancy is suppressed (filled circles). The corresponding onsite interaction strengths are $U/h = 0 \pm 80$ Hz and 5.0 ± 0.6 kHz, respectively. **b**, In the Mott insulating regime the double occupancy is strongly suppressed. It starts to increase for large atom numbers, indicating the formation of a metallic region in the trap centre. The blue and red lines represent the theoretical expectation for D in the atomic limit (see the text and Methods). Values and error bars are the means and s.d. of four to eight identical measurements. The systematic relative errors for the atom number, double occupancy and lattice depth are estimated to be 20%, 10% and 10%, respectively, with corresponding relative errors in J of up to 30%. These systematic errors apply to all further measurements.

2% for small atom numbers. This is direct evidence for the suppression of fluctuations in the occupation number and for the localization of the atoms.

To experimentally investigate the compressibility on entering the Mott insulating regime we determine how the double occupancy changes with increasing atom number; that is, we extract the slope $\partial D/\partial N$ from curves such as those shown in Fig. 2. This slope is a good measure of the compressibility $\kappa = \partial n/\partial \mu$ in those regions of the cloud where the filling n is near unity or larger, because n increases with D. We estimate the filling in the trap centre for the non-interacting case from the measured double occupancy[20]. It significantly exceeds one atom per site; for example, $\langle \hat{n} \rangle = 1.4$ for $N = 5 \times 10^4$, $V_0 = 7E_r$ and a temperature T of 30% of the Fermi temperature T_F.

The slope $\partial D/\partial N$ is shown in Fig. 3 for a wide range of interaction strengths. The data show that we access two regimes: for small onsite interaction energies U the slope $\partial D/\partial N$ is positive and the system is compressible, yet for $U/h > 5$ kHz the measured compressibility vanishes. This indicates that we have entered the Mott insulating regime. It implies a large central region with a filling reduced to one atom per site, surrounded by a metallic region with lower filling.

Further insight is gained by comparing our measurements with the theoretical values of $\partial D/\partial N$ calculated in the atomic limit[21] of the Hubbard model, including confinement and finite temperature. In this limit the kinetic energy is neglected by setting the tunnelling matrix element J to zero (see also Methods). We find good agreement between theory (black line in Fig. 3) and experimental data for $U \gg 6J$, where the above assumption is acceptable. For the calculation we have assumed a temperature of $T = 0.28T_F$, which is deduced from the entropy in the dipole trap as discussed in Methods. For zero temperature the slope $\partial D/\partial N$ would vanish as soon as U becomes larger than the chemical potential μ, which is $h \times 2.7$ kHz for $N = 8 \times 10^4$ atoms and a lattice potential of $V_0 = 12E_r$. Both our measurements and the model at finite temperature show a finite

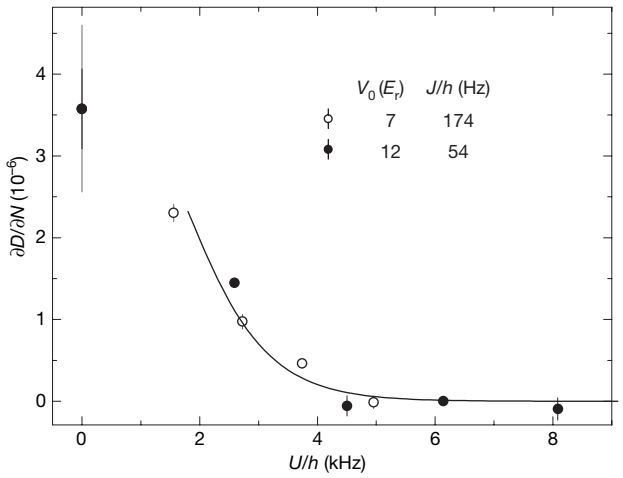


Figure 3 | The transition to an incompressible sample. On changing U, two regimes can be distinguished by the slope $\partial D/\partial N$. For vanishing interaction the large initial slope signals the filling of the Bloch band. Increasing U reduces the change in double occupancy. For $U/h \gtrsim 5$ kHz a change in atom number can no longer change the double occupancy. The compressibility $\partial D/\partial N$ is obtained from a least-squares fit of $D(N) = (\partial D/\partial N)N + D_0$ to data such as that shown in Fig. 2, with atom numbers in the interval from 25×10^3 to 8×10^4. Error bars denote the confidence interval of the fit. The expected slope in the atomic limit is illustrated with a black line for a lattice depth of $12E_r$ and $T = 0.28T_F$.

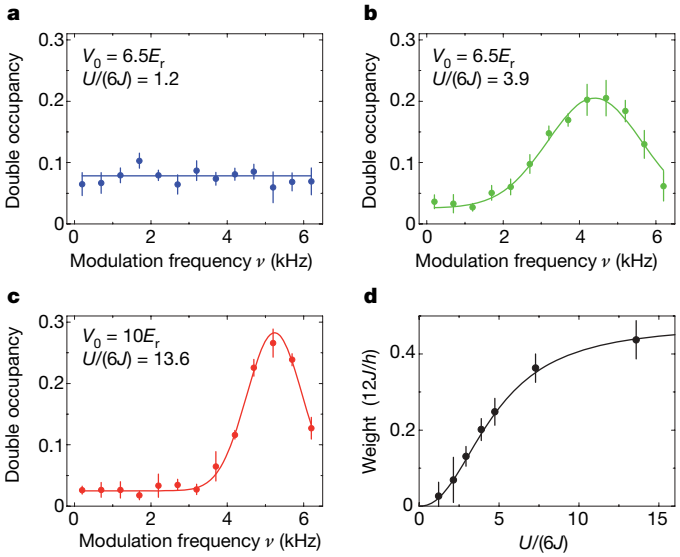


Figure 4 | Emergence of a gapped mode. **a**–**c**, With increasing interaction (blue (**a**), green (**b**), red (**c**)) the response to modulation of the lattice depth shows the appearance of a gapped mode. The weight of this peak grows with $U/(6J)$ and saturates. All modulation spectra were obtained with $(32 \pm 2) \times 10^3$ atoms. **d**, The weight of the peaks defined as

$$\sum_i \Delta\nu \left\{ D(\nu_i) - \frac{1}{2}[D(200\ \text{Hz}) + D(700\ \text{Hz})] \right\}$$

is shown. Here $D(\nu_i)$ is the measured double occupancy at frequencies ν_i, which are evenly spaced in steps of $\Delta\nu = 500$ Hz. It is plotted in units of $12J/h$. The first four data points are taken for a lattice depth of $6.5E_r$, the next at $7E_r$, $8E_r$ and $10E_r$, from left to right, respectively. The lines serve as a guide to the eye. Values and error bars are the mean and s.d. of four to eight identical measurements.

compressibility extending beyond $U/h = 2.7$ kHz, which can be attributed to thermal excitations. For the largest attained interaction $U/h = 8.1$ kHz, the thermal excitations are characterized by $T = 0.11U/k = 0.28T_F$, corresponding to 3% vacancies in the trap centre according to the theoretical analysis presented in Methods (k is Boltzmann's constant). The vanishing slope $\partial D/\partial N$ at this filling implies incompressibility of the core. The obtained ratio T/U is comparable to estimates for the bosonic Mott insulator[22].

In the strongly repulsive regime, the measured compressibility should vanish if $\mu < U$. For atom numbers corresponding to higher chemical potentials a metallic phase will appear in the trap centre and the double occupancy will increase. We observe this characteristic behaviour[23], which is a consequence of the presence of a Mott insulator (Fig. 2b). The behaviour agrees well with that expected from the Hubbard model in the atomic limit (lines in Fig. 2b). The free parameters in the theory curves, the temperature and a constant offset in D are determined by a least-squares fit to the data. The fits yield temperatures of $(0.2 \pm 0.1)T_F$. However, the accuracy is limited by the high sensitivity to the energy gap and the harmonic confinement. The constant offset in D accounts for the finite double occupancy in the ground state caused by second-order tunnelling processes as well as a systematic offset of 0.5% stemming from technical imperfections in the initial preparation of the spin mixture.

An important feature of a Mott insulator is the energy gap in the excitation spectrum. The lowest-lying excitations are particle–hole excitations centred at an energy U. The actual gap in the energy spectrum is reduced with respect to this value because of the width of the energy bands experienced by particles and holes[24]. A suitable technique for probing this excitation spectrum is to measure the response of the quantum gas in the optical lattice to a modulation of the lattice depth[25–27]: we apply 50 cycles of sinusoidal intensity modulation of all three lattice beams with an amplitude of 10%. The response is quantified by recording the double occupancy as a function of modulation frequency. With increasing interactions we observe the emergence of a gapped mode in the excitation spectrum (Fig. 4). For small values of $U/(6J)$, the double occupancy is not affected by the modulation of the lattice depth, but for large values of $U/(6J)$ a distinct peak appears for modulation frequencies ν near U/h. Furthermore, the area under the excitation curve divided by $12\ J/h$ as a measure for its width increases with interaction strength and starts to saturate at large values of $U/(6J)$ (see Fig. 4d).

The approach to the physics of the repulsive Fermi–Hubbard model that we have presented is completely different and complementary to that encountered in solid state systems, and it provides a new avenue to one of the predominant concepts in condensed matter physics. In this first experiment we have found clear evidence for the formation of a Mott insulator of fermionic atoms in the optical lattice. We could set limits for the deviation from unity filling in the Mott insulator by directly measuring the residual double occupancy and by deducing the number of holes from a realistic estimate of the temperature. The temperature is found to be small compared with the onsite interaction energy and the Fermi temperature. In addition, we have obtained good quantitative agreement with the Hubbard model in the atomic limit for a wide range of parameters. In further investigations of, for example, the energy spectra, the high resolution achieved may give direct insights into the width of Hubbard bands[24], the lifetime of excitations and the level of antiferromagnetic ordering[11,26] in the system.

METHODS SUMMARY

In the atomic limit $U \gg 6J$ of the Hubbard model we assume full localization of the fermions and thus neglect the kinetic energy. Each site is treated in the grand canonical ensemble with three possible occupation numbers $n \in \{0,1,2\}$. The partition function Z_i for site i is then

$$Z_i = \Sigma_n z^n \exp(-\beta E_{i,n}) = 1 + 2z\exp(-\beta\varepsilon_i) + z^2\exp(-2\beta\varepsilon_i - \beta U)$$

where $\beta = 1/kT$ is the inverse temperature, $z = \exp(\beta\mu)$ is the fugacity, μ is the chemical potential, $E_{i,n}$ is the energy of n particles on site i, and ε_i is the energy offset due to the harmonic confinement. For the probability of finding a double occupancy $\langle d\rangle_i$ or a vacancy $\langle v\rangle_i$ one obtains $\langle d\rangle_i = z^2\exp(-2\beta\varepsilon_i - \beta U)/Z_i$ and $\langle v\rangle_i = 1/Z_i$. Double occupancy D and total particle number N of the system are

obtained by summing over all sites; for example, $D = \Sigma_i 2\langle d\rangle_i / N$, where the equation for N is first solved numerically with respect to z. The entropy is

$$S = \frac{\partial}{\partial T}\left(kT \sum\nolimits_i \ln Z_i\right)$$

We calculate the temperature in the lattice by assuming that this entropy is the same as the entropy determined from temperature measurements in the dipole trap. The fits in Fig. 2b involve U as determined by modulation spectroscopy ($U/h = 4.7 \pm 0.1$ and 6.1 ± 0.1 kHz) because band structure calculations disagree with the measured value by up to 30% for the largest scattering lengths.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 15 April; accepted 9 July 2008.

1. Mott, N. F. *Metal–Insulator Transitions* (Taylor & Francis, 1990).
2. Imada, M., Fujimori, A. & Tokura, Y. Metal–insulator transitions. *Rev. Mod. Phys.* **70,** 1039–1263 (1998).
3. Lee, P. A., Nagaosa, N. & Wen, X.-G. Doping a Mott insulator: physics of high-temperature superconductivity. *Rev. Mod. Phys.* **78,** 17–85 (2006).
4. Hubbard, J. Electron correlations in narrow energy bands. *Proc. R. Soc. Lond. A* **276,** 238–257 (1963).
5. Jaksch, D., Bruder, C., Cirac, J. I., Gardiner, C. W. & Zoller, P. Cold bosonic atoms in optical lattices. *Phys. Rev. Lett.* **81,** 3108–3111 (1998).
6. Hofstetter, W., Cirac, J. I., Zoller, P., Demler, E. & Lukin, M. D. High-temperature superfluidity of fermionic atoms in optical lattices. *Phys. Rev. Lett.* **89,** 220407 (2002).
7. Greiner, M., Mandel, O., Esslinger, T., Hänsch, T. W. & Bloch, I. Quantum phase transition from a superfluid to a Mott insulator in a gas of ultracold atoms. *Nature* **415,** 39–44 (2002).
8. Trebst, S., Schollwöck, U., Troyer, M. & Zoller, P. *d*-wave resonating valence bond states of fermionic atoms in optical lattices. *Phys. Rev. Lett.* **96,** 250402 (2006).
9. Bloch, I., Dalibard, J. & Zwerger, W. Many-body physics with ultracold gases. *Rev. Mod. Phys.* **80,** 885–964 (2008).
10. Giorgini, L., Pitaevskii, L. P. & Stringari, S. Theory of ultracold Fermi gases. Preprint at ⟨http://arxiv.org/abs/0706.3360⟩ (2007).
11. Georges, A. in *Ultracold Fermi Gases* (eds Inguscio, M., Ketterle, W. & Salomon, C.) 477–533 (IOS Press, 2007).
12. Köhl, M., Moritz, H., Stöferle, T., Günter, K. & Esslinger, T. Fermionic atoms in a three dimensional optical lattice: observing Fermi surfaces, dynamics, and interactions. *Phys. Rev. Lett.* **94,** 080403 (2005).
13. Stöferle, T., Moritz, H., Günter, K., Köhl, M. & Esslinger, T. Molecules of fermionic atoms in an optical lattice. *Phys. Rev. Lett.* **96,** 030401 (2006).
14. Chin, J. K. *et al.* Evidence for superfluidity of ultracold fermions in an optical lattice. *Nature* **443,** 961–964 (2006).
15. Rom, T. *et al.* Free fermion antibunching in a degenerate atomic Fermi gas released from an optical lattice. *Nature* **444,** 733–736 (2006).
16. Strohmaier, N. *et al.* Interaction-controlled transport of an ultracold Fermi gas. *Phys. Rev. Lett.* **99,** 220601 (2007).
17. Helmes, R. W., Costi, T. A. & Rosch, A. Mott transition of fermionic atoms in a three-dimensional optical trap. *Phys. Rev. Lett.* **100,** 056403 (2008).
18. Rigol, M., Muramatsu, A., Batrouni, G. G. & Scalettar, R. T. Local quantum criticality in confined fermions on optical lattices. *Phys. Rev. Lett.* **91,** 130403 (2003).
19. Georges, A., Kotliar, G., Krauth, W. & Rozenberg, M. J. Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions. *Rev. Mod. Phys.* **68,** 13–125 (1996).
20. Köhl, M. Thermometry of fermionic atoms in an optical lattice. *Phys. Rev. A* **73,** 031601(R) (2006).
21. Gebhard, F. *The Mott Metal–Insulator Transition—Models and Methods* (Springer, 1997).
22. Gerbier, F. Boson Mott insulators at finite temperatures. *Phys. Rev. Lett.* **99,** 120405 (2007).
23. Gerbier, F., Fölling, S., Widera, A., Mandel, O. & Bloch, I. Probing number squeezing of ultracold atoms across the superfluid–Mott insulator transition. *Phys. Rev. Lett.* **96,** 090401 (2006).
24. Brinkman, W. F. & Rice, T. M. Single-particle excitations in magnetic insulators. *Phys. Rev. B* **2,** 1324–1338 (1970).
25. Stöferle, T., Moritz, H., Schori, C., Köhl, M. & Esslinger, T. Transition from a strongly interacting 1D superfluid to a Mott insulator. *Phys. Rev. Lett.* **92,** 130403 (2004).
26. Kollath, C., Iucci, A., McCulloch, I. P. & Giamarchi, T. Modulation spectroscopy with ultracold fermions in an optical lattice. *Phys. Rev. A* **74,** 041604(R) (2006).
27. Huber, S. D., Theiler, B., Altman, E. & Blatter, G. Amplitude mode in the quantum phase model. *Phys. Rev. Lett.* **100,** 050404 (2008).

Acknowledgements We thank J. Blatter, S. Huber, M. Köhl, C. Kollath, L. Pollet, N. Prokof'ev, M. Rigol, M. Sigrist, M. Troyer and W. Zwerger for discussions. Funding was provided by the Swiss National Science Foundation (SNF), the EU projects Optical Lattices and Quantum Information (OLAQUI) and Scalable Quantum Computing with Light and Atoms (SCALA) and the Quantum Science and Technology (QSIT) project of ETH Zurich.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to H.M. (moritz@phys.ethz.ch).

METHODS

Preparation. After sympathetic cooling with ^{87}Rb, 2×10^6 fermionic ^{40}K atoms are transferred into a dipole trap operating at a wavelength of 826 nm. Initially, a balanced spin mixture of atoms in the $|m_F\rangle = |-9/2\rangle$ and $|-7/2\rangle$ states is prepared and evaporatively cooled at a magnetic bias field of 203.06 G. Using this mixture we realize non-interacting samples with a scattering length of $a = (0 \pm 10)a_0$. Repulsive interactions are obtained by transferring the atoms in the $|-7/2\rangle$ state to the $|-5/2\rangle$ state during the evaporation, thus cooling and preparing a spin mixture of atoms in the $|-9/2\rangle$ and $|-5/2\rangle$ states, close to a Feshbach resonance at 224.21 G (ref. 28). After tuning the scattering length to the desired value, we load the atoms into the lowest Bloch band of the optical lattice by increasing the intensity of three retroreflected laser beams within 200 ms with the use of a spline ramp. The beams have circular profiles with $1/e^2$ radii of (160, 180, 160) μm at the position of the atoms. For a given scattering length and lattice depth, J and U are inferred from the Wannier functions including the interaction-induced coupling to the second Bloch band. The latter leads to corrections of up to 15% in U with respect to the single-band model.

Radiofrequency spectroscopy. By increasing the depth of the optical lattice to $30E_r$ in 0.5 ms, tunnelling is suppressed. The magnetic field is tuned to 201.28 G, where a molecular state for a $|-9/2\rangle$, $|-7/2\rangle$ pair with binding energy $h \times 99 \pm 1$ kHz and a weakly interacting state for a $|-9/2\rangle$, $|-5/2\rangle$ pair exist[13]. A radio-frequency π-pulse dissociates (associates) pairs and changes the spin state of those $|-7/2\rangle$ ($|-5/2\rangle$) atoms that share a site with a $|-9/2\rangle$ atom. Finally the magnetic field is increased to 202.80 G, dissociating any molecules, and the lattice potential is ramped down in 10 ms. All confining potentials are switched off and the homogeneous magnetic bias field is replaced by a magnetic gradient field in the same direction applied for 2 ms, thus spatially separating the spin states.

Imaging. After 6 ms of time-of-flight all three clouds are imaged simultaneously. As a result of a reproducible change of the imaging beam profile between the atomic absorption image and the subsequent reference image without atoms, residual structures are present in the density profiles. These are reduced by repeating the entire experiment without loading atoms and subtracting the obtained residual density distribution from the atomic density distribution. The number of atoms N_{m_F} per spin component m_F is determined from the two-dimensional column densities by simultaneously fitting the sum of three quartic terms $A \cdot \max(1 - (x/w_x)^4, 0) \cdot \max(1 - (y/w_y)^4, 0)$ with identical widths $w_{x,y}$ and mutual distances. This permits accurate detection of atom numbers down to 200 atoms per spin state. We have validated the absolute accuracy of the fits against integration of the density. The fraction D of atoms residing on doubly occupied sites is defined as $D = 2N_{m_{F'}}/N$, where $N = N_{-9/2} + N_{-7/2} + N_{-5/2}$ and $m_{F'} = -5/2$ $(-7/2)$ for samples initially containing atoms in the $|-7/2\rangle$ ($|-5/2\rangle$) state. The relative uncertainty in D is 10%, validated against measurements of the adiabatic molecule formation efficiency[13,16]. We estimate the relative systematic error for the total atom number N to be less than 20%. The $|-9/2\rangle$, $|-5/2\rangle$ mixture shows an offset of 0.5% in D as a result of $|-7/2\rangle$ atoms remaining from the initial spin transfer during evaporation.

Temperature. The temperature is measured in the harmonic dipole trap before ramping up the lattice and after a subsequent reversed ramp back into the dipole trap. The highest temperatures measured before and after ramping are $T_i = 0.15T_F$ and $T_f = 0.24T_F$, respectively. Because we expect non-adiabatic heating to occur during the lattice ramp up as well as during ramp down, we use the mean value of $0.195T_F$ as a realistic estimate. With this we calculate a temperature of $T = 0.28T_F$ in the Mott insulating regime ($a = 810a_0$, $V_0 = 12E_r$, $N = 10^5$), corresponding to 3.3% holes and a compressibility as low as $\partial n/\partial\mu = 0.09/\mu$ in the centre. For the temperatures in the dipole trap before and after the lattice ramp we would obtain 0.3% holes for T_i and 11.5% for T_f. The reported temperatures represent upper limits, because we have achieved temperatures down to $0.08T_F$ in the dipole trap before loading. Owing to inelastic collisions we lose at most (4.8 ± 0.6)% of the atoms during the preparation of the Mott insulating state for the parameters above, where the losses are expected to be highest. The inelastic decay time for atoms on doubly occupied sites exceeds 850 ms, which is significantly longer than the relevant experimental timescale.

28. Regal, C. A. & Jin, D. S. Measurement of positive and negative scattering lengths in a Fermi gas of atoms. *Phys. Rev. Lett.* **90**, 230404 (2003).

Vol 455|11 September 2008|doi:10.1038/nature07226

LETTERS

The transpiration of water at negative pressures in a synthetic tree

Tobias D. Wheeler[1] & Abraham D. Stroock[1]

Plant scientists believe that transpiration—the motion of water from the soil, through a vascular plant, and into the air—occurs by a passive, wicking mechanism. This mechanism is described by the cohesion-tension theory: loss of water by evaporation reduces the pressure of the liquid water within the leaf relative to atmospheric pressure; this reduced pressure pulls liquid water out of the soil and up the xylem to maintain hydration[1–3]. Strikingly, the absolute pressure of the water within the xylem is often negative, such that the liquid is under tension and is thermodynamically metastable with respect to the vapour phase[1,4]. Qualitatively, this mechanism is the same as that which drives fluid through the synthetic wicks that are key elements in technologies for heat transfer[5], fuel cells[6,7] and portable chemical systems[8–10]. Quantitatively, the differences in pressure generated in plants to drive flow can be more than a hundredfold larger than those generated in synthetic wicks. Here we present the design and operation of a microfluidic system formed in a synthetic hydrogel. This synthetic 'tree' captures the main attributes of transpiration in plants: transduction of subsaturation in the vapour phase of water into negative pressures in the liquid phase, stabilization and flow of liquid water at large negative pressures (−1.0 MPa or lower), continuous heat transfer with the evaporation of liquid water at negative pressure, and continuous extraction of liquid water from subsaturated sources. This development opens the opportunity for technological uses of water under tension and for new experimental studies of the liquid state of water.

To clarify the challenges of engineering a synthetic tree, Fig. 1a provides a minimal representation of the components required for the cohesion-tension mechanism of transpiration: a root membrane, a liquid-filled xylem capillary and a leaf membrane. The assembly in Fig. 1a acts as a passive conduit for water to travel from higher chemical potential in the soil, $\mu_{w,soil}$, to lower chemical potential in the air, $\mu_{w,air}$ (Fig. 1b, left). In general, both the water in the soil and in the air exist as subsaturated phases, for example as vapours with activities $a^{air}_{w,vap} \leq a^{soil}_{w,vap} \leq 1$ ($a_{w,vap} \equiv p_w/p_{w,sat}$ = (relative humidity)/100, where p_w and $p_{w,sat}$ are the actual and saturation vapour pressures). In contrast, the bulk liquid water in the xylem is nearly pure, with activity $a^{root}_{w,liq} = a^{leaf}_{w,liq} \approx 1$ ($a_{w,liq} \equiv$ effective mole fraction of water in solution[11]; Fig. 1b, centre). To accommodate these mismatched activities, the leaf and root membranes each mediate a local equilibrium in which the reduced activity in the external vapour is balanced by a reduction in the pressure in the internal liquid (Fig. 1b, right). We can solve for these differences in pressure in the balance of chemical potentials between the internal and external phases ($\mu_{w,soil} = \mu_{w,root}$ and $\mu_{w,leaf} = \mu_{w,air}$):

$$\Delta P^{root}_{mem} = P_{soil} - P_{root} = -\frac{RT}{\tilde{V}_{w,liq}} \ln\left(a^{soil}_{w,vap}\right) \text{ and}$$
$$\Delta P^{leaf}_{mem} = P_{air} - P_{leaf} = -\frac{RT}{\tilde{V}_{w,liq}} \ln\left(a^{air}_{w,vap}\right) \quad (1)$$

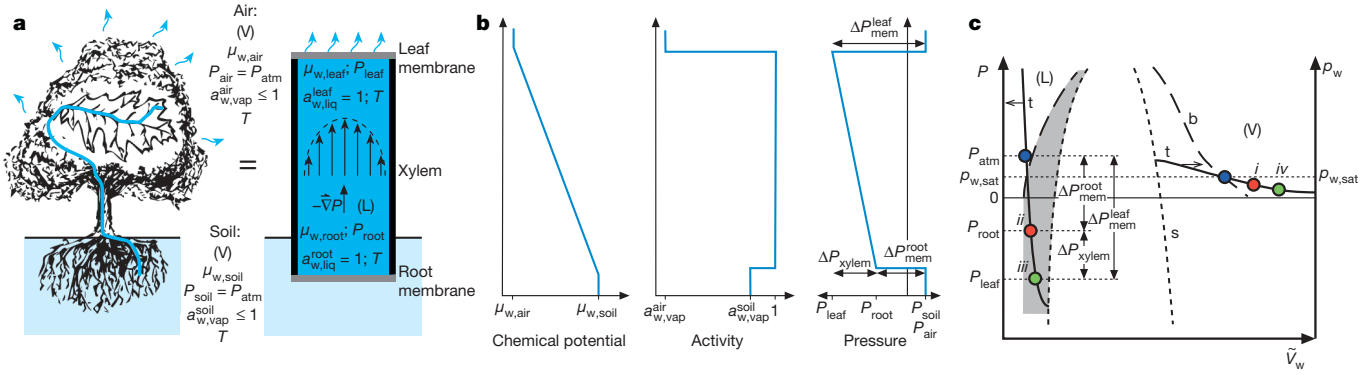


Figure 1 | Transpiration of water at negative pressures. **a**, Schematic representations of a tree through which water travels from the soil to the air (left) and of a minimal model of the components that support transpiration (right). At the root, water is drawn from the soil (pressure $P_{soil} = P_{atm}$; activity $a^{soil}_{w,vap} \leq 1$; temperature T) across a membrane and into the xylem. In this work, water in the soil is taken to be in the vapour phase (V). Within the xylem, liquid water (L) moves as a pressure-driven flow from root (P_{root}; $a^{root}_{w,liq} = 1$; T) to leaf ($P_{leaf} = P_{root} - \Delta P_{xylem}$; $a^{leaf}_{w,vap} = 1$; T). At the leaf, water in the xylem diffuses across a membrane and into the air ($P_{air} = P_{atm} > P_{leaf}$; $a^{air}_{w,vap} \leq 1$; T). **b**, Schematic profiles of chemical potential, activity and pressure of water as it passes through the tree. **c**, Schematic representation of an isotherm (t), the binodal (b) and the spinodal (s) in the projection of the phase diagram of water onto pressure (total pressure, P, for liquid (L) on left vertical axis and vapour pressure, p_w, for vapour (V) on right vertical axis) and molar volume, $\tilde{V}_w$. The pair of blue points indicates a stable liquid state at P_{atm} in equilibrium with a saturated vapour at $p_{w,sat}$. The pairs of red and green points indicate liquid states (ii and iii) in metastable equilibrium with subsaturated vapours (i and iv). During transpiration from a subsaturated soil to a subsaturated atmosphere, water follows the trajectory i → ii → iii → iv.

[1]School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, New York 14853, USA.

where R (J mol^{-1} K^{-1}) is the gas constant, $\tilde{V}_{w,liq}$ is the molar volume of liquid water ($\sim 1.8 \times 10^{-5}$ m^3 mol^{-1}), and the other parameters are defined in Fig. 1a. Important consequences of this membrane-mediated coupling are that, first, a difference exists in the pressure of water in the xylem between the root and the leaf, $\Delta P_{xylem} = P_{root} - P_{leaf} = \Delta P^{leaf}_{mem} - \Delta P^{root}_{mem} = (RT/\tilde{V}_{w,liq}) \ln\left(a^{soil}_{w,vap} / a^{air}_{w,vap}\right)$, when $P_{soil} = P_{air}$; this pressure difference drives the flow of water up the tree. Second, the pressure of water in the xylem is below atmospheric and can easily fall below zero (P_{root}, $P_{leaf} < 0$ for $a^{soil}_{w,vap}$ and $a^{air}_{w,vap} < 0.99925$). Pressures in the xylem of actual leaves have been measured[12] to be as low as $P_{leaf} = -10$ MPa. When P_{root}, $P_{leaf} < p_{w,sat}$ (shaded region in Fig. 1c), liquid water is mechanically stable ($\partial P/\partial V < 0$) and thermodynamically metastable with respect to the vapour state[13]. In other words, the liquid can act as a tensile element to pull itself up the tree, but it is prone to breakage by the formation of vapour embolisms (cavitation). These embolisms grow from pre-existing or spontaneously formed nuclei of vapour[13]. See the Supplementary Information for the derivation of equation (1) and for further comments on this minimal representation of transpiration.

Large negative pressures ($P \ll -1.0$ MPa) have been reported in static volumes of liquid water, as reviewed elsewhere[14]; the techniques used in those studies have not allowed for net flow. Hayward developed a mechanical pump that moved water up to a height of 17 m against gravity ($P = -0.07$ MPa)[15]. Despite the seeming simplicity of the mechanism by which plants manipulate liquids at significant negative pressures (Fig. 1), no one has reported the complete replication of transpiration in a synthetic system. The identification of a material to serve as the membranes in the root and leaf presents a central challenge to the development of such a system.

To address this challenge, we first consider a discrete volume of liquid separated from a subsaturated vapour by a water-permeable material (Fig. 2A). The conventional conception of this material, both in models of plant physiology[2,3,16] and designs of synthetic wicks[5,17,18], is of a micro- or nanoporous solid (Fig. 2A, top expanded view) in which the reduction of pressure within the pores by capillarity ($\Delta P_{cap} = P_{air} - P_{pore} = (2\sigma \cos\theta)/r_p$, where σ is the surface tension of water in J m^{-1}, pressure is in Pa and pore radius r_p is in m) allows for the phase equilibrium in equation (1): $\Delta P_{cap} = \Delta P_{mem}$. A number of groups have reported the generation of tension in bulk, static volumes of liquids coupled to vapours via nanoporous ceramics[2,16,17,19] (maximum tension $P = -1.2$ MPa in water[17]) and glass ($P = -4$ MPa in butane)[20]. Mechanical fragility[20] and heterogeneity in the sizes of pores[17,19] limited the magnitude of the tensions achieved and the practicality of these materials.

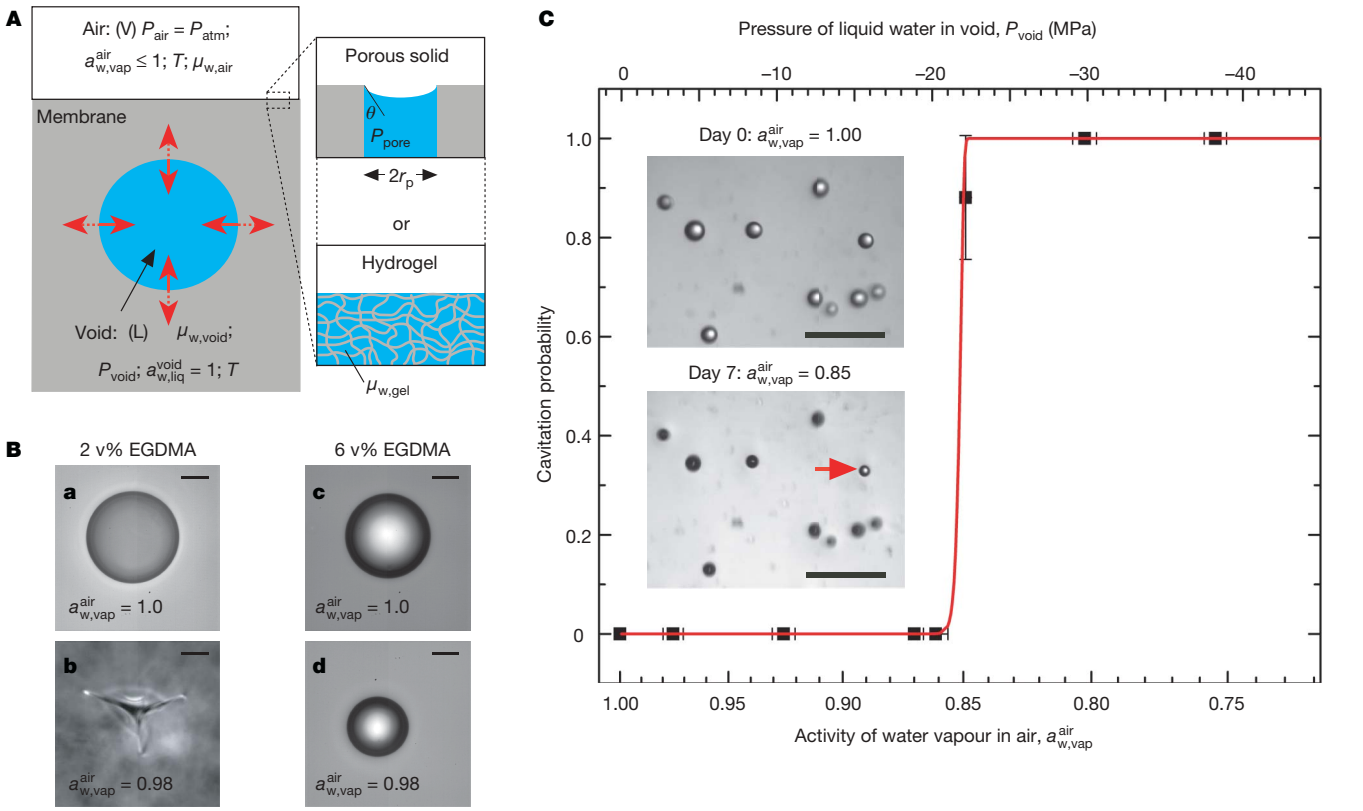


Figure 2 | Liquid water in equilibrium with subsaturated vapours. **A**, Schematic cross-sectional view of liquid water entrapped within a spherical void within a membrane. Pure liquid in the void ($a^{void}_{w,liq} = 1$) equilibrates across the membrane with water vapour in the air ($a^{air}_{w,vap} \leq 1$). At equilibrium, the pressure within the liquid $P_{void} = P_{atm} - \Delta P^{void}_{mem}$, according to equation (1). This stress (solid red arrows) is negative and is balanced by the mechanical response of the membrane (dashed red arrows). The expanded view indicates the composition of the membrane: either a porous solid with pores of radius r_p (m), filled with water at pressure P_{pore} (Pa) and contact angle θ (deg), or a hydrogel containing water of chemical potential $\mu_{w,gel}$ (J mol^{-1}). **B**, Optical micrographs of water-filled voids in two hydrogel membranes (pHEMA) of differing concentration of cross-linker (EGDMA), at equilibrium with $a^{air}_{w,vap} = 1$ (**a** and **c**) and with $a^{air}_{w,vap} = 0.98$ (**b** and **d**). Scale bars, 50 μm. **C**, Plot of the probability of cavitation of the liquid water within voids within pHEMA (6 vol% EGDMA) after equilibration with vapours of different $a^{air}_{w,vap}$. The pressure within the liquid, P_{void}, calculated with equation (1) for each activity, is plotted on the top axis. For each point, the state of $n \geq 100$ voids was followed and the experiment was repeated three times. Vertical error bars are the standard deviation of the probability of cavitation over the three repetitions. Horizontal error bars are the uncertainty in the $a^{air}_{w,vap}$ generated by the saturated salt solutions used (see Supplementary Information). The red curve is a fit of the data to a kinetic theory of cavitation (Supplementary Information, equation (8)) with a threshold pressure $P_{cav} = -21.3$ MPa. Inset micrographs show example populations of voids. Scale bars, 500 μm. After equilibration with $a^{air}_{w,vap} = 0.849 \pm 0.005$, most voids were empty (dark). The red arrow indicates a liquid-filled void.

As an alternative, we have investigated chemically cross-linked, organic hydrogels as membranes (Fig. 2A, bottom expanded view); experiments with non-cross-linked hydrogels[19] set a precedent for this choice of material. Water within a hydrogel acts as a solution with regard to both its thermodynamics[21] and dynamics[22]. As such, hydrogels can maintain both hydration (avoid invasion of air) and the mobility of water as they equilibrate ($\mu_{w,air} = \mu_{w,gel} = \mu_{w,void}$) with vapours of arbitrarily low activity ($a^{air}_{w,vap} \rightarrow 0$)[21,22]; this behaviour is in contrast to the invasion of air and precipitous drop in permeability to water that occurs in rigid, porous solids at non-zero activities[23].

To evaluate hydrogels as membranes, we created liquid-filled voids directly within sheets of poly(hydroxyethyl methacrylate) (pHEMA). We allowed these sheets to equilibrate with vapours of well-defined activity (Fig. 2A; see Methods Summary). The micrographs in Fig. 2B show liquid-filled voids in pHEMA in equilibrium with $a^{air}_{w,vap} = 1.0$ (Fig. 2Ba, c) and with $a^{air}_{w,vap} = 0.98$ (Fig. 2Bb, d). The collapse of the void in a soft hydrogel formed with low concentrations of chemical cross-linker (Fig. 2Ba, b) provides a qualitative demonstration of the reduced pressure ($P_{void} = P_{atm} - \Delta P^{void}_{mem} = -3.3$ MPa via equation (1)) in the entrapped liquid on equilibration with a subsaturated vapour. In stiffer hydrogels formed with higher concentrations of cross-linker (Fig. 2Bc, d), the void resisted collapse.

To establish the limit of stability of the liquid state within these voids, we measured the probability of cavitation of liquid water equilibrated with vapours of various activities (Fig. 2C). The probability undergoes a sharp transition at $a^{air}_{w,vap} = 0.849 \pm 0.005$. This transition

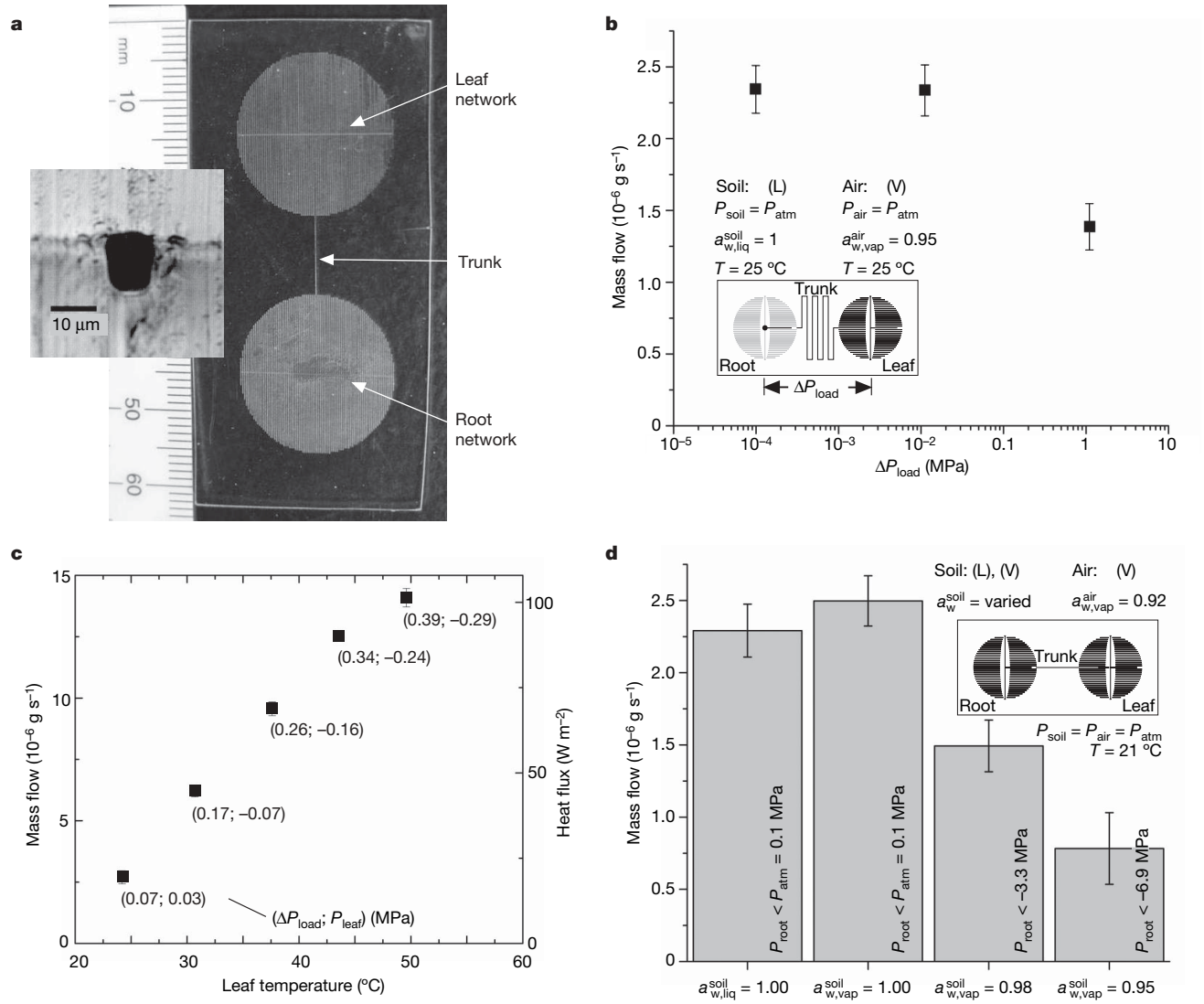


Figure 3 | Transpiration through a synthetic tree. a, Optical photograph of a synthetic tree: a transparent sheet of pHEMA (1 mm thick) containing a void in the form of a microchannel network at its mid-plane. The structures of the networks in the root and leaf are identical: 80 parallel channels of varying length arranged to form a circle and connected by a single orthogonal channel. The inset shows an optical micrograph of the cross-section of one microchannel (dark area). **b**, **c**, Mass flow rate of water driven by transpiration from a reservoir of pure liquid water above the root to a stream of air with $a^{air}_{w,vap} = 0.95$ above the leaf (see inset in **b**). The root membrane was punctured at the centre of the root network. In **b**, mass flows for three different trees are shown; each tree had a trunk channel with a different hydraulic load resistance, depending on the effective hydraulic diameter, D_E, and the length, L, of the trunk channel (low: $D_E = 73.5$ μm, $L = 3$ cm; intermediate: $D_E = 22.5$ μm, $L = 3$ cm; high: $D_E = 11.6$ μm, $L = 35$ cm) (see Supplementary Information). Vertical error bars represent the standard deviation of flow rates measured over time. Horizontal error bars representing the uncertainty in calculated values of ΔP_{load}, obtained through the propagation of the uncertainty in measurements of flow, are smaller than the data points. In **c**, mass flow rate of water and heat flux are plotted against the temperature maintained beneath the leaf section. The pressure drop across the trunk, ΔP_{load}, and the pressure at the entrance of the leaf network, $P_{leaf} = P_{atm} - \Delta P_{load}$, are reported next to each data point. The trunk dimensions in this experiment were $D_E = 14.5$ μm and $L = 3$ cm. Vertical error bars are for flow, as in **b**. **d**, Mass flow rate by transpiration through a synthetic tree with intact root membrane and with liquid (L) and vapours (V) of various activities, a^{soil}_w, above the root. Vertical error bars are for flow as in **b**.

corresponds to a predicted pressure in the liquid water (equation (1)), $P_{void} = -22.15 \pm 00.72$ MPa, more than 10 times the largest tension reported for a wick[17]. The observed probability of cavitation as a function of pressure is consistent with the prediction of a kinetic model of nucleation (red trend line in Fig. 2C) proposed by Herbert *et al.*[24] (see Supplementary Information); this agreement suggests that the breakage of the liquid phase is due to thermally activated nucleation of the vapour phase in the bulk of the liquid or at its boundaries, and not due to invasion of air through the membrane or the expansion of pre-existing nuclei.

We can use the membrane-mediated coupling of liquid water to subsaturated vapours to drive the motion of liquid water at negative pressures. To achieve this function, we used soft lithography to create voids in sheets of pHEMA in the form of a microfluidic network; we call this device the synthetic tree (Fig. 3a; see Methods Summary). We achieved steady-state flow through the synthetic tree by exposing the leaf to a stream of air with $a^{air}_{w,vap} < 1$ and exposing the root to pure liquid water ($a^{soil}_{w,liq} = 1$) or a stream of air with $a^{soil}_{w,vap} > a^{air}_{w,vap}$ (see Methods Summary).

The leaf section of the synthetic tree can act alone as a pump to pull liquid through hydraulic loads. Figure 3b presents the flow rates of water pulled through three loads as a function of the drop in pressure across the load, ΔP_{load}. To simplify the evaluation of the hydraulic resistances, we perforated the root membrane. For the highest load, $\Delta P_{load} = 1.1$ MPa and $P_{leaf} = P_{atm} - \Delta P_{load} = -1.0$ MPa; this negative pressure is 15-fold greater in magnitude than the highest value previously reported for a synthetic pumping system[15]. At higher loads (not shown), the liquid in the synthetic tree cavitated before a steady state was achieved; cavitation in these systems may have been induced at lower tensions than in the spherical voids (Fig. 2C) owing to the entry of contaminants or vapour nuclei provided by the flow through the hole in the root membrane.

Evaporation from the leaf of a synthetic tree can serve to cool a heat source. In this context, the capacity of the leaf to act as a tension-based pump allows it to maintain continuous evaporative cooling with a source of liquid that is separated from the site of evaporation by a large hydraulic load or gravitational pressure head. Figure 3c presents mass flow and evaporative heat flux as a function of the temperature at the leaf of a synthetic tree operated as in Fig. 3b (intermediate load) with the addition of a heater below the leaf section. As expected, the mass flow grew as the temperature of the leaf rose; this behaviour offers one means of controlling the flux from a pump based on a synthetic tree. For temperatures above 30 °C, the pressure drop separating the evaporator (leaf) from the reservoir of liquid (soil) was $\Delta P_{load} > P_{atm}$, such that the liquid in the leaf was negative ($P_{leaf} < 0$). At 50 °C, which was the maximum temperature achieved without cavitation of the liquid in the channels, the pressure was $P_{leaf} = -0.29$ MPa; this tension could be used to move water tens of metres vertically against gravity to the evaporator.

A complete synthetic tree, with both leaf and root membranes intact (Fig. 3d, inset), can draw liquid from a reservoir that contains a subsaturated vapour ($a^{soil}_{w,vap} < 1$), as first suggested by Dixon and Joly more than 100 years ago[2]. In this case, the liquid in the microchannels transfers a portion of the reduced pressure generated in the leaf to the root, where it pulls water across the root membrane; the process in the root is analogous to reverse osmosis in which pure water is extracted from a dilute phase, in this case a subsaturated vapour in air. Figure 3d presents the flow rate of water drawn through a synthetic tree when the leaf was exposed to vapour with $a^{air}_{w,vap} = 0.92$ and the root was exposed to liquid water and to vapours with $a^{soil}_{w,vap} = 1.0$, 0.98 and 0.95. For each case, we report the upper bound on P_{root} based on equation (1) and the requirement that $P_{root} < \left(P_{soil} - \Delta P^{root}_{mem}\right)$, with $P_{soil} = P_{atm}$, to extract liquid water from the soil reservoir. For $a^{soil}_{w,vap} = 0.95$, the pressure in the root water $P_{root} < -6.9$ MPa. The stability of the liquid at this large tension (relative to the maximum achieved in synthetic trees with punctured root membranes; Fig. 3b, c) suggests that drawing water through a hydrogel membrane eliminates impurities that nucleate cavitation.

The realization in a synthetic system of the key features of the cohesion-tension mechanism of transpiration provides new support for this theory. The experiments we present demonstrate proof-of-principle for technologies that could use transpiration at negative pressures (1) to allow processes that demand large pressure differences (for example high-performance liquid chromatography) to be performed passively in microfluidic systems; (2) to extend greatly the maximum dimensions and heat flux of wick-based heat pipes[5]; and (3) to offer new approaches to extract and purify water from subsaturated soils. The effectiveness of hydrogel membranes in the synthetic tree suggests that similar materials may play analogous roles in plants; interestingly, hydrogels present in xylem capillaries participate in the regulation of flow[25]. Synthetic trees also provide an experimental platform with which to investigate other aspects of plant physiology (for example embolism recovery[26]) and fundamental properties of liquids (for example the origin of the thermodynamic and dynamic anomalies in water[27]).

METHODS SUMMARY

The pHEMA hydrogels were formed from a solution of 2-hydroxyethyl methacrylate (65 vol%), ethyleneglycol dimethacrylate (2 or 6 vol%), methacrylic acid (1 vol%), de-ionized water (32 or 28 vol%) and a photoinitiator, 2,2-dimethoxy-2-phenylacetophenone dissolved in n-vinyl pyrrolidone at 600 mg ml^{-1} (1 vol%). This solution was injected into a casting jig to generate 1-mm-thick sheets. Polymerization was initiated with ultraviolet light (385 nm). To create spherical voids within the hydrogel sheets, bubbles of air were introduced into the hydrogel solution before injection into the jig. To create synthetic trees, a polydimethylsiloxane (PDMS) stamp was defined using soft lithography[28]. The microfluidic network was transferred to a sheet of hydrogel (500 μm thick) by partial polymerization of the hydrogel precursor solution on the PDMS stamp. A second, flat sheet of polymerized hydrogel (500 μm thick) was placed over the network and the two sheets were bonded through further exposure to ultraviolet light. Spherical voids and networks were filled by soaking in water at 21 °C or 100 °C, or by pressurization in water to ~50 MPa. For cavitation experiments, the activity of water vapour was controlled using saturated salt solutions. For transpiration experiments, the activity of water vapour above leaf section was maintained with an impinging airstream that was conditioned using a dew-point generator. The flow of water through synthetic trees was determined by measuring the mass of a reservoir from which the trees pulled water. Connections between the trees and the reservoir were established using Plexiglas holders and high-performance liquid chromatography tubing and connections. Pressure drop across the trunk channel was calculated from the measured flow rates, trunk dimensions and Poiseuille's law. Trunk lengths were taken as the distance between the centres of root and leaf networks. Trunk cross-sections were measured from optical micrographs.

Received 10 February; accepted 30 June 2008.

1. Boehm, J. Capillaritat und Saftsteigen. *Ber. Dtsch. Bot. Ges.* **11,** 203–212 (1893).
2. Dixon, H. H. & Joly, J. On the ascent of sap. *Phil. Trans. R. Soc. Lond. B* **186,** 563–576 (1895).
3. Nobel, P. S. *Physicochemical and Environmental Plant Physiology* 2nd edn (Academic, 1999).
4. Scholander, P. F., Hammel, H. T., Bradstreet, E. D. & Hemmingsen, E. A. Sap pressure in vascular plants: Negative hydrostatic pressure can be measured in plants. *Science* **148,** 339–346 (1965).
5. Peterson, G. P. *An Introduction to Heat Pipes: Modeling, Testing, and Applications* (Wiley, 1994).
6. Chen, J., Matsuura, T. & Hori, M. Novel gas diffusion layer with water management function for PEMFC. *J. Power Sources* **131,** 155–161 (2004).
7. Karan, K. *et al.* An experimental investigation of water transport in PEMFCs: The role of microporous layers. *Electrochem. Solid State Lett.* **10,** B34–B38 (2007).
8. Effenhauser, C. S., Harttig, H. & Kramer, P. An evaporation-based disposable micropump concept for continuous monitoring applications. *Biomed. Microdevices* **4,** 27–32 (2002).
9. Guan, Y. X., Xu, Z. R., Dai, J. & Fang, Z. L. The use of a micropump based on capillary and evaporation effects in a microfluidic flow injection chemiluminescence system. *Talanta* **68,** 1384–1389 (2006).
10. Juncker, D. *et al.* Autonomous microfluidic capillary system. *Anal. Chem.* **74,** 6139–6144 (2002).
11. Atkins, P. & de Paula, J. *Physical Chemistry* 7th edn (Oxford Univ. Press, 2002).
12. Jacobsen, A. L., Pratt, R. B., Ewers, F. W. & Davis, S. D. Cavitation resistance among 26 chaparral species of southern California. *Ecol. Monogr.* **77,** 99–115 (2007).

13. Debenedetti, P. G. *Metastable Liquids* (Princeton Univ. Press, 1996).
14. Caupin, F. & Herbert, E. Cavitation in water: A review. *C. R. Phys.* **7,** 1000–1007 (2006).
15. Hayward, A. T. J. Mechanical pump with a suction lift of 17 metres. *Nature* **225,** 376–377 (1970).
16. Askenasy, E. Beitrage zur Erklarung des Saftsteigens. *Verh. Naturhist. Med. Ver. Heidelberg* **5,** 429–448 (1896).
17. Guan, Y. & Fredlund, D. G. Use of the tensile strength of water for the direct measurement of high soil suction. *Can. Geotech. J.* **34,** 604–614 (1997).
18. Aybar, H. S., Egelioglu, F. & Atikol, U. An experimental study on an inclined solar water distillation system. *Desalination* **180,** 285–289 (2005).
19. Thut, H. F. Demonstration of the lifting power of evaporation. *Ohio J. Sci.* **28,** 292–298 (1928).
20. Machin, W. D. A simple method for the generation of negative pressure in liquids. *Can. J. Chem. Rev. Can. Chim.* **76,** 1578–1580 (1998).
21. Flory, P. J. *Principles of Polymer Chemistry* (Cornell Univ. Press, 1955).
22. Arce, A., Fornasiero, F., Rodriguez, O., Radke, C. J. & Prausnitz, J. M. Sorption and transport of water vapor in thin polymer films at 35 degrees C. *Phys. Chem. Chem. Phys.* **6,** 103–108 (2004).
23. Brinker, C. J & Scherer, G. W. *Sol-gel Science: The Physics and Chemistry of Sol-Gel Processing* (Academic, 1989).
24. Herbert, E., Balibar, S. & Caupin, F. Cavitation pressure in water. *Phys. Rev. E* **74,** 041603 (2006).
25. Zwieniecki, M. A., Melcher, P. J. & Holbrook, N. M. Hydrogel control of xylem hydraulic resistance in plants. *Science* **291,** 1059–1062 (2001).
26. Holbrook, N. M. & Zwieniecki, M. A. Embolism repair and xylem tension: Do we need a miracle? *Plant Physiol.* **120,** 7–10 (1999).
27. Errington, J. R. & Debenedetti, P. G. Relationship between structural order and the anomalies of liquid water. *Nature* **409,** 318–321 (2001).
28. McDonald, J. C. & Whitesides, G. M. Poly(dimethylsiloxane) as a material for fabricating microfluidic devices. *Acc. Chem. Res.* **35,** 491–499 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. A. Zwieniecki, N. M. Holbrook, P. J. Melcher, C. Cohen, K. J. Niklas, C. Cottin-Bizonne, A. Lassiter and F. Caupin for discussions and suggestions. We thank G. Swan and E. Velez-Rosa for technical assistance with experiments. Support was provided by the Office of Naval Research Young Investigator Program and the Camille and Henry Dreyfus Foundation. T.D.W. acknowledges partial support by a graduate fellowship from the Corning Foundation. The experiments made use of the following facilities: Cornell NanoScale Science and Technology Facility (a member of the National Nanotechnology Infrastructure Network, supported by the National Science Foundation (NSF)), the Nanobiotechnology Center (supported by the STC Program of the NSF under Agreement No. ECS-98767710) and the Cornell Center for Materials Research (supported by the NSF under Award No. DMR-0520404).

Author Contributions A.D.S. conceived the project. Both authors designed the experiments. T.D.W. executed the experiments. Both authors wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.D.S. (ads10@cornell.edu).

LETTERS

Upward migration of Vesuvius magma chamber over the past 20,000 years

B. Scaillet[1,2], M. Pichavant[1,2] & R. Cioni[3,4]

Forecasting future eruptions of Vesuvius is an important challenge for volcanologists, as its reawakening could threaten the lives of 700,000 people living near the volcano[1,2]. Critical to the evaluation of hazards associated with the next eruption is the estimation of the depth of the magma reservoir, one of the main parameters controlling magma properties and eruptive style. Petrological studies have indicated that during past activity, magma chambers were at depths between 3 and 16 km (refs 3–7). Geophysical surveys have imaged some levels of seismic attenuation, the shallowest of which lies at 8–9 km depth[8], and these have been tentatively interpreted as levels of preferential magma accumulation. By using experimental phase equilibria, carried out on material from four main explosive events at Vesuvius, we show here that the reservoirs that fed the eruptive activity migrated from 7–8 km to 3–4 km depth between the AD 79 (Pompeii) and AD 472 (Pollena) events. If data from the Pomici di Base event 18.5 kyr ago[9] and the 1944 Vesuvius eruption[7] are included, the total upward migration of the reservoir amounts to 9–11 km. The change of preferential magma ponding levels in the upper crust can be attributed to differences in the volatile content and buoyancy of ascending magmas, as well as to changes in local stress field following either caldera formation[10] or volcano spreading[11]. Reservoir migration, and the possible influence on feeding rates[12], should be integrated into the parameters used for defining expected eruptive scenarios at Vesuvius.

The comparison of natural and experimental phase assemblages obtained from hydrothermal experiments represents a powerful approach to assess pre-eruption conditions, in particular pressure, in subvolcanic magma reservoirs, as shown for the eruptions of Mount St Helens[13], Mount Pelée[14] and Mount Pinatubo[15], among others. To this end, we investigated the phase equilibria of phonolite magma from four main explosive events of Vesuvius: Mercato (7800 BP), Avellino (3600 BP), Pompeii (79 AD) and Pollena (472 AD). Rocks selected for the experiments are representative of the first material erupted during each Plinian event (see Supplementary Information). This magma has been interpreted[16–20] as being derived from the apical part of the reservoir, being the most evolved of the erupted products. Although magma mixing affects most of the products erupted during large eruptions[16–19], petrological and geochemical studies suggest that the first phonolites to be erupted did not experience this process. The compositional and mineralogical differences of the four studied phonolites can be interpreted in terms of different conditions of magma evolution and crystallization. In particular, isotopic analyses show that in the studied phonolites, phenocrysts and matrix glass share the same $^{87}Sr/^{86}Sr$ composition[17,19], suggesting that the phonolitic cap of the reservoir approached crystal–liquid equilibrium before eruption.

We used established laboratory procedures to carry out crystallization experiments (see Methods) to work out the phase relationships of the four phonolites over the pressure and temperature ranges 100–300 MPa and 750–900 °C, and with varying fluid composition. Based on 216 experiments, both isobaric/polythermal and isothermal/polybaric sections were drawn. The isobaric/isothermal sections of Pompeii and Pollena phonolites are shown in Fig. 1. The full set of experimental data for each eruption can be found in the Supplementary Information. Phase relationships suggest similar equilibration pressures ($P \approx 200$ MPa) for the three older phonolites (Mercato, Avellino, Pompeii). This is in general agreement with data from melt inclusions[17] and geochemical studies[21], which suggest a process of recycling of magma residuals in the same reservoir for the three eruptions. Conversely, our experiments document an important difference in pre-eruption pressure between the Pompeii and Pollena events.

Inspection of Fig. 1 shows that the phenocryst assemblage of Pompeii phonolite, consisting of sanidine, amphibole, garnet, clinopyroxene, leucite (rare), and plagioclase, can be reproduced at 200 MPa under a restricted range of conditions of temperature and dissolved water content in the melt (H_2O_{melt}), broadly centred at 815 ± 10 °C and 6 wt% dissolved H_2O. The polybaric phase relations established at 800 °C (see Supplementary Information) show in contrast that pressures significantly lower than 200 MPa do not reproduce the observed phase assemblage, in particular because amphibole is no longer stable near the liquidus. Altogether, a pressure of 200 ± 20 MPa is taken as the condition prevailing in the upper part of the zoned reservoir tapped by the Pompeii eruption. Phase equilibrium considerations for Mercato and Avellino phonolites lead to a similar estimate for the pressure of magma storage (Table 1, see Supplementary Information). In contrast, at 200 MPa the phenocryst assemblage of Pollena phonolite cannot be reproduced experimentally. Clinopyroxene would be largely present at subliquidus conditions at 200 MPa, but clinopyroxene and nepheline cannot coprecipitate at near-liquidus conditions, as demanded by the phenocryst assemblage and the crystal-poor character of Pollena pumice. The only way to crystallize nepheline and clinopyroxene together near the liquidus is to decrease pressure down to 100 MPa in the experiments (Fig. 1b). This decrease in pressure promotes leucite growth at the expense of sanidine, in agreement with the petrographic features of Pollena pumice, which shows rare sanidine and abundant leucite (Supplementary Information, and ref. 20), whereas the reverse is observed in Pompeii phonolite. As a general fact, leucite is nearly absent in pre-Pompeii magmas, but abundant in post-Pompeii deposits[22].

Additional evidence for such a difference in pressure between the Pompeii and Pollena reservoirs comes from the study of volatiles trapped in melt inclusions[6,23]. Pompeii pumice is characterized by pre-eruptive dissolved water contents of 6–7 wt% (ref. 23), in good agreement with the range predicted from phase equilibrium

[1]CNRS/INSU—Institut des Sciences de la Terre d'Orléans, 1a rue de la Férollerie, 45071 Orléans, cedex 2, France. [2]Université d'Orléans—Institut des Sciences de la Terre d'Orléans, 45071 Orléans, France. [3]Dipartimento di Scienze della Terra, Via Trentino 51, 09127 Cagliari, Italy. [4]INGV, sezione di Pisa, Via della Faggiola, 56100 Pisa, Italy.

considerations (Fig. 1). In contrast, the water content of melt inclusions in Pollena phonolite phenocrysts (3–5 wt%; ref. 6) indicates trapping pressures around 100 MPa (ref. 6). We conclude that the level of preferential magma accumulation migrated upward between the Pompeii and Pollena events. Considering an average crustal density of 2,600 kg m^{-3}, such a difference of 100 MPa would correspond to a depth difference of about 4 km.

Pre-eruption conditions are less well constrained for events either older than the Mercato or younger than the Pollena eruptions, although they can be obtained from field and petrological data. The Pomici di Base eruption, the first Plinian event at Somma-Vesuvius, dated around 18.5 kyr ago[9], ejected trachytic to latitic magmas. A thermobarometric approach using feldspar phenocrysts from this eruption[9] has yielded equilibrium pressure in the range 300–500 MPa. Although these estimates are associated with a significant uncertainty, they clearly point towards pressure of magma storage equivalent to, or even higher than, that of younger Plinian events. Since the sub-Plinian Pollena event, a number of eruptions have occurred, including the 1631 event, which has considerable similarities in eruption style, magma composition and mineral assemblage to the Pollena event[24,25]. As in the Pollena rocks, the widespread occurrence of leucite phenocrysts in the 1631 tephra strongly suggests that low pressure conditions prevailed during pre-eruptive magma storage, possibly at around 100 MPa. Following the 1631 eruption, the activity of Vesuvius shifted towards a semi-persistent state, which continued up to the last event in 1944. This period was characterized by a Strombolian-type dynamism with alternating explosive and effusive phases[26,27], with intervening repose times not longer than 7 yr and violent Strombolian eruptions as the most intense events[26,27]. In two of the largest eruptions of this period, in 1906 and 1944[5,7,26], melt inclusions in olivine and diopside testify to the arrival of deep mafic magma batches (pressures >300 MPa; refs 5,7) before the eruption. Salitic clinopyroxene and leucite phenocrysts host melt inclusions with H_2O contents hardly exceeding 1 wt% and no CO_2, yielding entrapment pressures in the range 50–80 MPa (ref. 7) which suggest that crystallization occurred in a reservoir no deeper than 3 km.

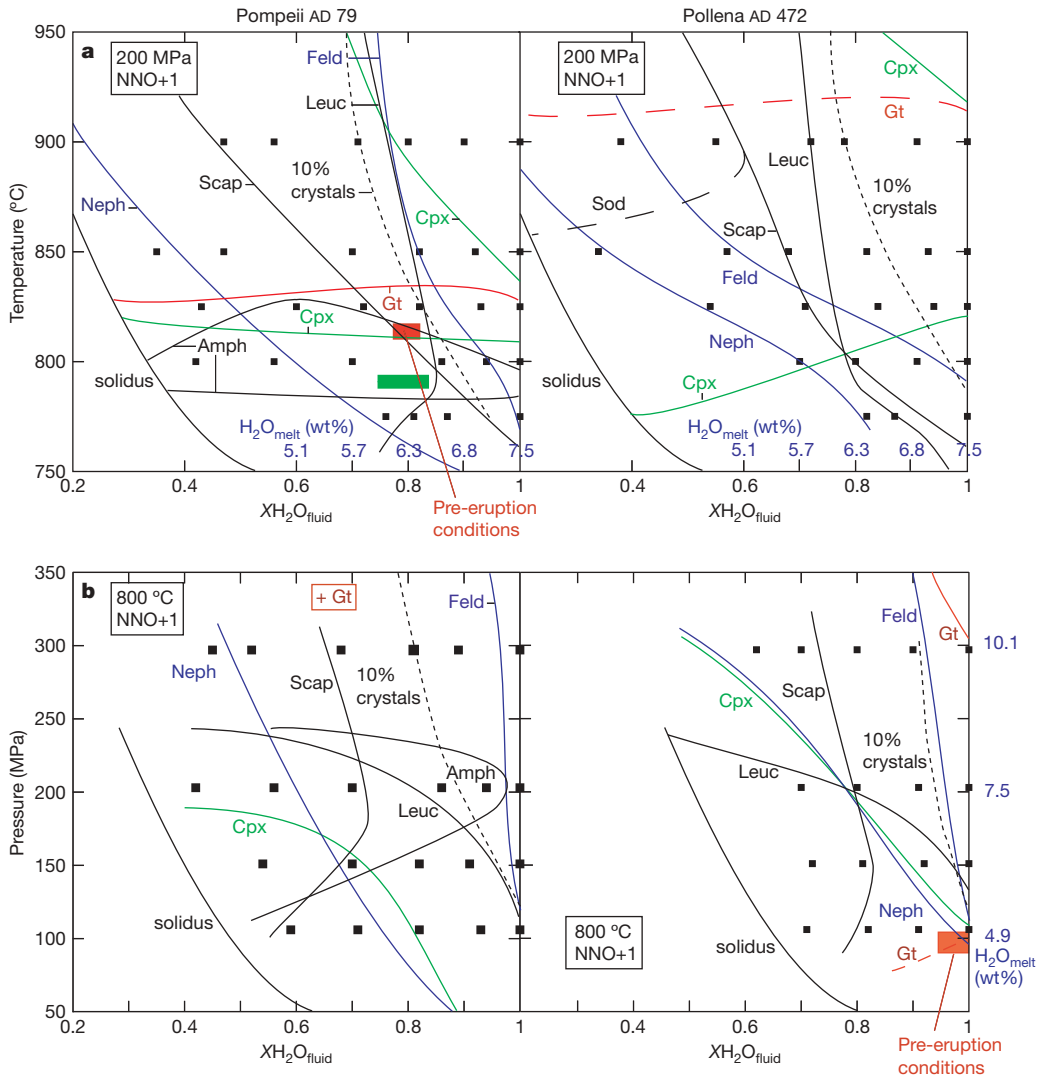


Figure 1 | Phase relationships of Pompeii and Pollena phonolites. **a**, Isobaric/polythermal phase relationships of Pompeii and Pollena phonolites at 200 MPa. Phases lie inside their stability field. The dashed line gives the 10% modal proportion of crystals in run products. The red box shows the estimated pre-eruption conditions of temperature and XH_2O_{fluid} (XH_2O_{fluid} being the mole fraction of H_2O in the fluid phase), as inferred from the phenocryst assemblage of the Pompeii phonolite. The correspondence between XH_2O_{fluid} and H_2O_{melt} is given along the lower horizontal axis. The thick green bar is the range of H_2O_{melt} obtained from the analysis of melt inclusions in sanidine phenocrysts of the Pompeii phonolite[23]. For Pollena, note the impossibility of co-crystallizing clinopyroxene and nepheline while maintaining crystal-poor conditions (10%). **b**, Isothermal/polybaric phase relationships of Pompeii and Pollena phonolites at 800 °C. The water contents at saturation are given along the vertical right axis. Leucite is not stable at pressures much higher than 200 MPa, notably under H_2O-rich conditions. The red box shows the estimated pre-eruption T–XH_2O conditions of Pollena, as inferred from the phenocryst assemblage of the Pollena phonolite. Phase abbreviations are: Feld: feldspar; Leuc: leucite; Amph: amphibole; Gt: garnet; Cpx: clinopyroxene; Neph: nepheline; Scap: scapolite; Sod: sodalite. Note that in both compositions, biotite is present in runs below 850 °C, and plagioclase occurs along with sanidine, although the latter largely predominates over the former (see Supplementary Information). Black squares represent the T–XH_2O_{fluid} location of individual experimental charges.

Table 1 | Storage conditions of Vesuvius magmas over the past 20 kyr

Eruption	Age (kyr)	*P* (MPa)	*T* (°C)	H_2O_{melt} (wt%)	XH_2O_{fl}	MgO (wt%)
Pomici di Base	18.5	300–500	830–860	>5	–	0.45
Mercato	8.01	200 ± 20	785 ± 10	7 ± 0.5	0.95	0.13
Avellino	3.36	200 ± 20	785 ± 10	5–5.5	0.65	0.22
Pompeii	1.92	200 ± 20	815 ± 10	6–6.5	0.82	0.42
Pollena	1.53	100 ± 20	800 ± 10	3–4	>0.95	0.7
1631	0.377	100(?)	977 ± 30	–	–	1.7
1906	0.102	50–80	>1,100	0.8–1.2	–	2.9
1944	0.062	50–80	1,100	0.8–1.1	–	2.6

Sources: P–T–H_2O_{melt} conditions for Mercato, Avellino, Pompeii and Pollena events are from this work. Other data are from ref. 9 for the Pomici di Base, and refs 7,18,26 for the 1906 and 1944 eruptions. The pre-eruption temperature for the 1631 eruption was calculated using the CaO content of the most felsic groundmass glass[25] and CaO geothermometry[18]. The pre-eruption temperature of the 1944 eruption was also calculated using the CaO content[18] of groundmass glass[7]. For all eruptions, P–T–H_2O_{melt} conditions correspond to the topmost part of the crustal reservoir. XH_2O_{fl} is the mole fraction of H_2O in the coexisting fluid phase.

Figure 2a plots the pressure of magma storage and crystallization for selected Vesuvius eruptions (Table 1) against eruption age. Clearly, there is a decrease in pressure from 300 ± 100 MPa at 18.5 kyr ago, down to less than 100 MPa for the 1906–44 events. Thus, our study highlights the existence of various, yet not synchronous, shallow reservoirs beneath Vesuvius in which chemical differentiation occurred. Petrological and geochemical arguments show that these shallow reservoirs grew over time by repeated injection of mafic batches[17–19] coming from a deeper reservoir or directly from the source region[5–7]. In addition to this age–depth correlation, the available data show that there is a close relationship between extent of magma differentiation, repose time and depth of the reservoir (Fig. 2b). Reservoirs of Plinian eruptions, all located at pressures of at least 200 MPa, invariably produced trachyte to phonolite magmas with MgO contents lower than 0.5 wt% (Table 1) and erupted after a period of quiescence ranging from hundreds to thousands of years[18,22]. The shallower reservoirs, which characterized the recent Pollena and 1631 events, fractionated up to tephriphonolitic compositions, being able to produce only small volumes of phonolite (MgO between 0.7 and 1.7 wt%) (Table 1). In both cases, the production of felsic derivatives testifies to the possibility of significant magma cooling and differentiation during upper crustal residence. In contrast, the magma erupted during the 1944 and 1906 eruptions is a potassium-rich tephrite to phonotephrite with MgO contents always higher than 2.5 wt% (Table 1). Inversion of diffusion profiles in 1944 clinopyroxenes shows that residence of deep phenocrysts within the shallow reservoir was 9 yr or less[28]. Altogether, this suggests that the reservoir feeding those recent eruptions, besides being extremely shallow, has a short lifetime, as mafic magma feeding the shallower system was possibly erupted soon after its injection[26,27]. The great variability in magma types and eruptive styles documented at Vesuvius therefore appears to be closely correlated with variations in depths of magma storage, pre-eruptive temperature and volatile contents. Overall, the above lines of evidence show that the shallower the reservoir, the hotter, drier and more mafic is the average composition of the erupted magma (Fig. 2).

The strong control of temperature and volatile content over eruptive regimes, and their tight relationship with pressure, imply that forecasting the next eruption at Vesuvius requires us to know the present level of magma ponding. The fact that past phonolitic to trachytic reservoirs lay at a level similar to the shallowest seismic attenuation zone revealed by seismic tomography (Fig. 2a) could mean that the seismic discontinuity reflects a present-day growing phonolitic reservoir. The subsurface levels of magma ponding may be related to the presence of lithological or structural discontinuities, such as the upper boundary of the carbonate basement[27], which may either halt or aid the ascent of magma batches. In addition, upward migration of the level of magma ponding, whether progressive or step-like, could imply that the magma properties or the plumbing system, or both, have experienced an irreversible change over time. Melt inclusion studies[17] have revealed a change in the major element composition of the feeding magmas, from K-basaltic to tephritic, and this is mirrored by the evolution trends of erupted rocks[18,22], in particular during the period between the Avellino and Pollena eruptions. This may imply a difference in volatile content, and thus buoyancy, of the feeding magmas. Along the same lines, recent work has suggested the possibility of extensive carbonate ingestion of mafic magmas at Vesuvius, with massive production of CO_2 which may also affect magma buoyancy[29]. Change in the local stress field[10] following caldera formation[22,24], or gravitational spreading of the volcano onto its weak sedimentary substratum[11], may also have played a part. Modelling work has shown that changes in shape and size of a volcanic edifice strongly affect eruption behaviour[12]. In particular, reduction of overburden and tension allows magma batches to pond at higher levels, with a consequent reduction of repose times and extent of differentiation.

Changes in pressure of magma storage similar to those observed at Vesuvius, again with the more felsic products associated with the

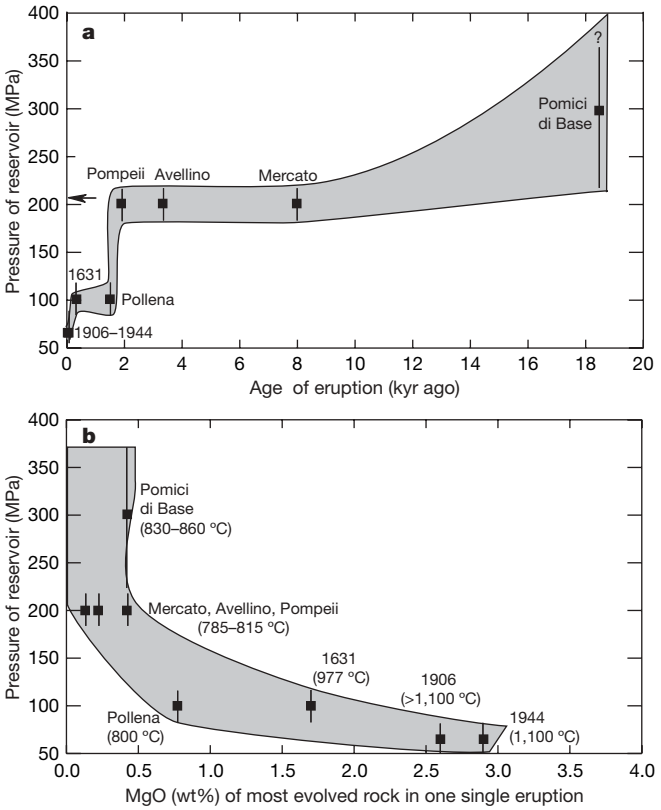


Figure 2 | Time evolution of the pressure of magma storage at Vesuvius for the past 20 kyr, as inferred from petrological constraints. a, We show only eruptions for which the petrology of erupted products has been studied in some detail. The grey field represents the time evolution of the pressure of magma storage. The horizontal arrow on the left vertical axis represents the approximate location of shallowest level of seismic attenuation identified by tomography[8]. **b**, Variation of the extent of chemical differentiation with the pressure of magma storage (grey field). We take as a differentiation index the bulk MgO content of the rock. Because the main fractionating phase assemblage is Cpx ± Ol, the MgO content of the residual liquid always decreases with fractionation. For each eruption we have reported the most evolved rock composition belonging to the corresponding volcanic deposit. Also indicated are the corresponding temperatures of the magma, assumed to represent the topmost part of the emptied reservoir (Table 1). Note that, for those eruptions that produced phonolitic magma, the general decrease in temperature is also mirrored by a decrease in the abundance of felsic derivatives in the erupted products. It varies from near 100%, as in the Mercato event, to a few per cent in the case of Pollena, down to 0% for all eruptions since 1631. The error bars represent the uncertainty in the pressure of magma storage as inferred either from experiments (this work) or petrological analysis (ref. 9, 7, 18, 26).

deeper reservoirs, have been documented at Mount St Helens (160 to 300 MPa; ref. 13), albeit occurring over a shorter timescale (4 kyr). An increase of the reservoir pressure acts to reduce the flux of magma from the deep source[12], assisting in protracted cooling and differentiation; such a feedback mechanism may have been operative at both Mount St Helens and Vesuvius. Although the weighted average rate of magma supply is estimated[22] to have remained broadly constant over the past 10,000 yr, variations in feeding rates over three orders of magnitude, closely associated to variations in eruption styles, have been documented during the 1631–44 period[27]. Therefore the outstanding question is whether such changes may have happened during periods preceding important explosive events.

Our findings have several implications for hazard mitigation at Vesuvius. First, numerical models aimed at simulating the dynamics of the next eruptions at Vesuvius have so far considered the shallow reservoir to have a fixed position in the upper crust[30], whereas our results show that it may have varied considerably over time. Second, models predicting the size of the next likely event are all based on the assumption that the magma feeding rates of upper reservoirs have remained broadly constant over at least the past 4 kyr. In the light of our results, the possibility that the feeding rate has fluctuated, in particular before Plinian events, should be considered. Third, it is of utmost importance that the type of fluid and/or magma stored at 8–9 km depth is properly identified. Geophysical methods cannot distinguish mafic from felsic magmas, being able to detect at best the presence of fluids encapsulated in solids. Substantial effort is needed to increase the chemical resolution of geophysical surveys.

METHODS SUMMARY

We investigated the phase equilibria of the four selected phonolites through crystallization experiments. We first melted the phonolites at 1 bar and 1,500 °C to obtain dry and crystal-free glasses, which were then ground fine. We prepared experimental charges by loading into gold capsules the powdered dry glass together with known amounts of water and CO_2 (added as silver oxalate). Capsules were welded shut before annealing at high P and T. Experiments were mostly done using an internally heated vessel, with Ar–H_2 mixtures as a pressurizing gas, and maintaining redox conditions between those of the nickel–nickel oxide buffer (NNO) and 1 log f_{O2} unit above this (NNO + 1). A few low-temperature experiments were also done in cold-seal pressure vessels, pressurized with argon. Redox conditions were continuously monitored with a semi-permeable H_2 membrane placed next to the sample holder. The thermal gradient across the sample holder was monitored with three to four inconel-sheathed thermocouples and was always less than 2 °C. Altogether, run temperatures and pressures are known to within ±5 °C and 2 MPa. Most experiments consisted of running the four phonolite compositions together, exploring for each a range of H_2O/CO_2 fluid ratios. Typically an experiment consisted of annealing 20 to 30 charges simultaneously, minimizing any variations in phase assemblages and compositions that might have arisen from differences in applied P–T–f_{H2}. Run durations varied according to temperature, between 4 and 28 days. Experiments were ended by isobaric quenching. After completion of the experiment, capsules were weighed to check for leaks, then opened. Half of the recovered run product was mounted in epoxy resin and polished for subsequent optical inspections, scanning electron microscopy and electron microprobe analyses.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 23 April; accepted 4 July 2008.

1. Guidoboni, E. & Boschi, E. Vesuvius before the 1631 eruption. *Eos* **87,** 417–423 (2006).
2. Heiken, G. Will Vesuvius erupt? Three million people need to know. *Science* **286,** 1685–1687 (1999).
3. Barberi, F. *et al.* The Somma-Vesuvius magma chamber: a petrological and volcanological approach. *Bull. Volcanol.* **44,** 295–315 (1981).
4. Belkin, H. E., De Vivo, B., Roedder, E. & Cortini, M. Fluid inclusion geobarometry from ejected Mt Somma-Vesuvius nodules. *Am. Mineral.* **70,** 288–303 (1985).
5. Marianelli, P., Sbrana, A., Métrich, N. & Cecchetti, A. The deep feeding system of Vesuvius involved in the recent violent strombolian eruptions. *Geophys. Res. Lett.* **32,** doi:10.1029/2004GL021667 (2005).
6. Fulignati, P. & Marianelli, P. Tracing volatile exsolution within the 472 AD 'Pollena' magma chamber of Vesuvius (Italy) from melt inclusion investigation. *J. Volcanol. Geotherm. Res.* **161,** 289–302 (2007).
7. Fulignati, P., Marianelli, P., Métrich, N., Santacroce, R. & Sbrana, A. Towards a reconstruction of the magmatic feeding system of the 1944 eruption of Mt Vesuvius. *J. Volcanol. Geotherm. Res.* **133,** 13–22 (2003).
8. Auger, E., Gasparini, P., Virieux, J. & Zollo, A. Seismic evidence of an extended magmatic sill under Mt Vesuvius. *Science* **294,** 1510–1512 (2001).
9. Landi, P., Bertagnini, A. & Rosi, M. Chemical zoning and crystallisation mechanisms in the magma chamber of the Pomici di Base plinian eruption. *Contrib. Mineral. Petrol.* **135,** 179–197 (1999).
10. Ventura, G., Vilardo, G. & Bruno, P. P. The role of flank failure in modifying the shallow plumbing system of volcano: an example from Somma-Vesuvius, Italy. *Geophys. Res. Lett.* **26,** 3681–3684 (1999).
11. Borgia, A. *et al.* Volcanic spreading of Vesuvius, a new paradigm for interpreting its volcanic activity. *Geophys. Res. Lett.* **32,** doi:10.1029/2004GL022155 (2005).
12. Pinel, V. & Jaupart, C. Magma chamber behavior beneath a volcanic edifice. *J. Geophys. Res.* **108,** doi:10.1029/2002JB001751 (2003).
13. Gardner, J. E., Rutherford, M., Carey, S. & Sigurdsson, H. Experimental constraints on pre-eruptive water contents and changing storage prior to explosive eruptions of Mount St Helens volcano. *Bull. Volcanol.* **57,** 1–17 (1995).
14. Martel, C. *et al.* Magma storage conditions and control of eruption regime in silicic volcanoes: experimental evidence from Mt Pelée. *Earth Planet. Sci. Lett.* **156,** 89–99 (1998).
15. Scaillet, B. & Evans, B. W. The June 15, 1991 eruption of Mount Pinatubo. I. Phase equilibria and pre-eruption P–T–fO_2–fH_2O conditions of the dacite magma. *J. Petrol.* **40,** 381–411 (1999).
16. Sigurdsson, H., Cornell, W. & Carey, S. Influence of magma withdrawal on compositional gradients during the AD 79 Vesuvius eruption. *Nature* **345,** 519–521 (1990).
17. Cioni, R. *et al.* Compositional layering and syn-eruptive mixing of a periodically refilled shallow magma chamber: the AD 79 Plinian eruption of Vesuvius. *J. Petrol.* **36,** 739–776 (1995).
18. Cioni, R., Marianellia, P. & Santacroce, R. Thermal and compositional evolution of the shallow magma chambers of Vesuvius: evidence from pyroxene phenocrysts and melt inclusions. *J. Geophys. Res.* **103,** 18277–18294 (1998).
19. Civetta, L. & Santacroce, R. Steady state magma supply in the last 3400 years of Vesuvius activity. *Acta Vulcanol.* **2,** 147–159 (1992).
20. Rosi, M. & Santacroce, R. The AD 472 Pollena eruption: volcanological and petrological data for this poorly known plinian-type event at Vesuvius. *J. Volcanol. Geotherm. Res.* **17,** 249–271 (1983).
21. Civetta, L., Galati, R. & Santacroce, R. Magma mixing and convective compositional layering within the Vesuvius magma chamber. *Bull. Volcanol.* **53,** 287–300 (1991).
22. Santacroce, R. (ed.) *Somma-Vesuvius* (CNR Quaderni de La Ricerca Scientifica, 1987).
23. Cioni, R. Volatile content and degassing processes in the AD 79 magma chamber at Vesuvius (Italy). *Contrib. Mineral. Petrol.* **140,** 40–54 (2000).
24. Rosi, M., Principe, C. & Vecci, R. The 1631 Vesuvius eruption. A reconstruction based on historical and stratigraphical data. *J. Volcanol. Geotherm. Res.* **58,** 151–182 (1993).
25. Rolandi, G., Barrella, A. M. & Borrelli, A. The 1631 eruption of Vesuvius. *J. Volcanol. Geotherm. Res.* **58,** 183–201 (1993).
26. Santacroce, R., Bertagnini, A., Civetta, L., Landi, P. & Sbrana, A. Eruptive dynamics and petrogenetic processes in a very shallow magma reservoir: the 1906 eruption of Vesuvius. *J. Petrol.* **34,** 383–425 (1993).
27. Scandone, R., Giacomelli, L. & Speranza, F. F. Persistent activity and violent strombolian eruptions at Vesuvius between 1631 and 1944. *J. Volcanol. Geotherm. Res.* **170,** 167–180 (2008).
28. Morgan, D. J. *et al.* Time scales of crystal residence and magma chamber volume from modelling of diffusion profiles in phenocrysts: Vesuvius 1944. *Earth Planet. Sci. Lett.* **222,** 933–946 (2004).
29. Marziano, I., Gaillard, G. & Pichavant, M. Limestone assimilation by basaltic magmas: an experimental re-assessment and application to Italian volcanoes. *Contrib. Mineral. Petrol.*. doi:10.1007/s00410–007–02677–8 (2008).
30. Todesco, M. *et al.* Pyroclastic flow hazard assessment at Vesuvius (Italy) by using numerical modelling. I. Large scale dynamics. *Bull. Volcanol.* **64,** 155–177 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Scandone for a review, and A. Sbrana, P. Marianelli and R. Santacroce for discussions. This project was financially supported by GNV and INGV funds and by the department of the Italian Civil Defense.

Author Contributions All authors participated to the definition of the overall project strategy and to the field campaign aimed at sample selection and collection. B.S. performed the experiments, analysed the run products and produced the first draft of the paper which other authors then discussed.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to B.S. (bscaille@cnrs-orleans.fr).

METHODS

Starting material and capsule preparation. For each phonolite, several hand-sized blocks of pumice were ground in an agate mortar down to a mesh size of 200–100 μm. Approximately 10 g of powdered pumice was loaded into a platinum crucible and melted at 1,500 °C for three hours in air, then ground, and remelted for three hours at 1,500 °C in air. The resulting dry glass was then ground down to a mesh size of 20 μm, and this powder was used as a starting material for all phase equilibrium experiments. The starting glass compositions were checked by electron microprobe analyses (see Supplementary Table 1) and compare well with published bulk rock analyses of pumices from similar stratigraphic levels, showing in particular that the alkali contents of the glasses do not differ significantly from that of the original rock. Similarly, the iron content of both glasses and rock powders are identical within analytical error, showing that iron loss towards the platinum crucible did not occur during atmospheric melting.

Gold capsules were used (inner diameter 2.5 mm, thickness 0.2 mm, length 15 mm), welded with a graphite arc welder. Distilled H_2O was first loaded, then silver oxalate as the source of CO_2 for H_2O-undersaturated runs, and then the glass powder. After welding, the capsules were left in an oven for a few hours at 100 °C, to ensure homogeneous H_2O distribution. We maintained the total amount of $H_2O + CO_2$ constant added to the capsule (3 ± 0.5 mg) and also kept constant the fluid/silicate ratio (30 ± 3 mg of silicate). At any given T–P conditions and for each phonolite composition, various starting H_2O–CO_2 mixtures were explored, with X_{H2Ofl} defined as $H_2O/(H_2O + CO_2)$ (in moles) varying in the range 1–0.2. Weighing before and after the run allowed us to check for capsule leaks at high pressure and temperature, and capsules showing differences in weight of less than 0.3 mg were considered as successful.

Experimental and analytical techniques. Experiments were mostly done with an internally heated vessel operating vertically, loaded with Ar–H_2 mixtures as the pressurizing gas, and fitted with an H_2-membrane[31] to monitor continuously the H_2 fugacity (f_{H2}). A run at 750 °C was made using a cold-seal pressure vessel, pressurized with pure argon which imposes redox conditions at NNO + 3 (3 log units above the NNO solid buffer). Each experiment contained up to 28 capsules located within the hot spot, with a thermal gradient of less than 2 °C across, the capsules sitting on top of the membrane. Runs performed with an H_2 membrane were isobarically quenched by switching off the power supply while maintaining the pressure within ±20 bar of the target value down to a temperature of 300 °C, which results in an overall cooling rate of 100 °C per minute in the temperature range 900–500 °C. The run at 750 °C was drop-quenched with a cooling rate of 100 °C s^{-1}. Run durations varied according to temperature between 4 and 28 days. The estimated uncertainties in temperature, total pressure and f_{H2} are ±5 °C, ±2 MPa and 0.01 MPa, respectively.

After the experiments, capsules were checked for leaks, then opened, and half of the run product, usually as a single fragment, was embedded in a probe mount with an epoxy resin and polished for optical observation, and for electron microprobe and scanning electron microscopy analyses. All charges were systematically observed by scanning electron microscopy, and the different phases present determined by semiquantitative energy-dispersive spectroscopy. Electron microprobe analysis was also used to check for consistency in selected charges, using analytical conditions as in ref. 15 (see Supplementary Table 2).

We made a total of 216 phase equilibrium experiments. The general strategy was to establish first a complete isobaric section at 200 MPa within the temperature range 750–900 °C, and then vary pressure at a temperature close to the pre-eruptive one, as inferred from isobaric experiments. We thus have established complete isothermal sections at 800 °C in the pressure range 100–300 MPa, but also at 775 °C in the pressure range 150–250 MPa. For each P–T and bulk composition a range of melt water content was explored. Although the phase relationships are shown in projections of T or P–XH_2O_{fl}, we calculated the melt water contents of charges bracketing the most likely pre-eruption conditions by using available solubility models for phonolites[32,33] and the following simple relationship:

$$f_{H2O,charge} = f_{H2O}{}^{0} \times XH_2O_{fl}$$

where $f_{H2O}{}^{0}$ is the fugacity of pure water at P and T, and XH_2O_{fl} the mole fraction of H_2O in the coexisting fluid phase.

31. Scaillet, B., Pichavant, M., Roux, J., Humbert, G. & Lefevre, A. Improvements of the Shaw membrane technique for measurement and control of f_{H2} at high temperatures and pressures. *Am. Mineral.* **77**, 647–655 (1992).
32. Carroll, M. & Blank, J. The solubility of H_2O in phonolitic melts. *Am. Mineral.* **82**, 549–556 (1997).
33. Iacono Marziano, G. & Schmidt, B. C. &. D. o. l. f. i. D. Equilibrium and disequilibrium degassing of a phonolitic melt (Vesuvius AD 79 ''White Pumice'') simulated by decompression experiments. *J. Volcanol. Geotherm. Res.* **161**, 151–164 (2007).

LETTERS

Understanding the limits to generalizability of experimental evolutionary models

Samantha E. Forde[1]*, Robert E. Beardmore[2]*, Ivana Gudelj[2,3]*, Sinan S. Arkin[2], John N. Thompson[1] & Laurence D. Hurst[4]

Given the difficulty of testing evolutionary and ecological theory *in situ, in vitro* model systems are attractive alternatives[1]; however, can we appraise whether an experimental result is particular to the *in vitro* model, and, if so, characterize the systems likely to behave differently and understand why? Here we examine these issues using the relationship between phenotypic diversity and resource input in the T7–*Escherichia coli* co-evolving system as a case history. We establish a mathematical model of this interaction, framed as one instance of a super-class of host–parasite co-evolutionary models, and show that it captures experimental results. By tuning this model, we then ask how diversity as a function of resource input could behave for alternative co-evolving partners (for example, *E. coli* with lambda bacteriophages). In contrast to populations lacking bacteriophages, variation in diversity with differences in resources is always found for co-evolving populations, supporting the geographic mosaic theory of co-evolution[2]. The form of this variation is not, however, universal. Details of infectivity are pivotal: in T7–*E. coli* with a modified gene-for-gene interaction, diversity is low at high resource input, whereas, for matching-allele interactions, maximal diversity is found at high resource input. A combination of *in vitro* systems and appropriately configured mathematical models is an effective means to isolate results particular to the *in vitro* system, to characterize systems likely to behave differently and to understand the biology underpinning those alternatives.

We start by considering a mathematical model tailored to the specific biology of the bacterium *E. coli* and its bacteriophage T7/T3 (for brevity, we refer to T7; we might equally be modelling T3, ref. 3) but framed in such a way that alternative co-evolving systems might also be analysed. The model tracks evolution in initially isogenic populations of co-occurring clonally reproducing bacteria (B_0) and phages (P_0) in the chemostat. Mutation occurs with a small but prescribed probability, and the fitness of mutant bacteria and phages depend on every component of the system (formally, a genotype-by-genotype-by-environment interaction, also known as a selection mosaic[2,4]).

Although host and parasites, including bacteria and phages[5], are often thought to interact along a continuum from 'gene-for-gene' to 'matching alleles' [6,7] (see Fig. 1a, b), neither of these models matches the known biology of the *E. coli*–T7 interaction because T7 phages have higher adsorption rates to wild-type *E. coli* than to contemporary hosts[8–10]. Thus, in our initial model, we suppose that the binding probabilities between bacteria and phages are graded (see ref. 11; Fig. 1c).

The graded infection mechanism is understandable when the details of the biology are known. Relative resistance to these phages can be conferred through mutations that truncate lipopolysaccharides (LPSs) found within the outer membrane, thus preventing adsorption of the phage[10,12]. These truncations can be shallow or deep[10,13]. Relative resistance is also conditioned by pleoitropic interactions between LPS and outer membrane proteins (OMPs) and is dependent on mutations in both (Supplementary Information 6.3). For simplicity, we assume that there are two character states at two loci, L and O, in wild-type bacteria (B_0). To capture the pleiotropy, we assume that mutations at these loci regulate the biosynthesis of LPS polymers such that the length of the LPS O antigen correlates with the phenotype $l = 4 - (2 \times L + O)$, yielding four phenotypes: $B_0(L = 0, O = 0)$ with $l = 4$; $B_1(0,1)$ with $l = 3$; $B_2(1,0)$ with $l = 2$; and $B_3(1,1)$ with $l = 1$. We term the graded mechanism (Fig. 1c) a 'modified gene-for-gene interaction' owing to its resemblance to gene-for-gene interactions (Fig. 1a).

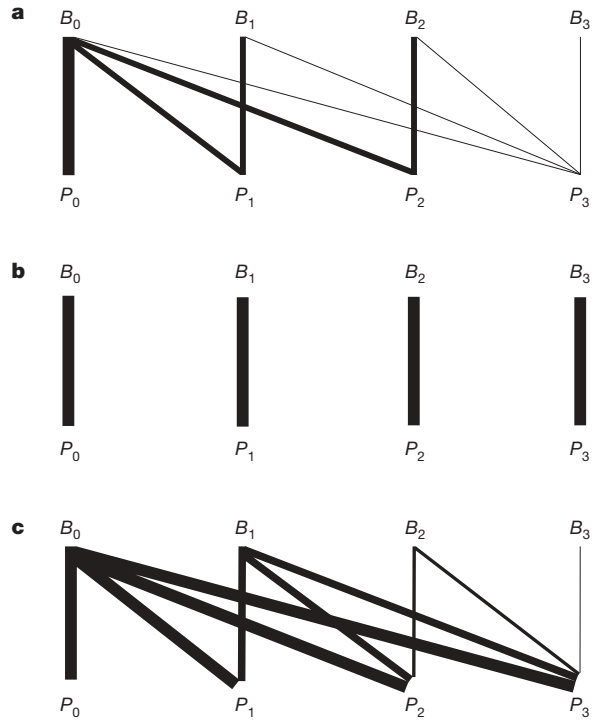


Figure 1 | Infection mechanisms between bacteria (*B*) and phages (*P*). **a–c**, The thickness of the lines represents infectivity levels: gene-for-gene with intrinsic cost of virulence[6] (**a**); matching alleles (**b**); modified gene-for-gene where infectivity is always highest on the ancestral host (**c**).

[1]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, California 95064, USA. [2]Department of Mathematics, Imperial College London, London SW7 2AZ, UK. [3]Department of Mathematical Sciences and [4]Department of Biology & Biochemistry, University of Bath, Bath BA2 7AY, UK.
*These authors contributed equally to this work.

Assumptions about pleiotropy determine the form of the mutational matrix M_b between the four bacterial types. We have considered numerous matrices and find that results are strikingly insensitive to M_b and M_p (Supplementary Information 6.3). Mutations in wild-type phage (P_0) occur on one locus with four possible alleles giving rise to one of three types, denoted P_i (where i is from 1 to 3).

The core of the *E. coli*–T7 model is a 4×4 matrix (Φ) that defines the relative infectivities of each phage strain to each bacterial type:

$$\Phi = \begin{array}{c} \begin{matrix} P_0 & P_1 & P_2 & P_3 \end{matrix} \\ \begin{pmatrix} 1 & \lambda & \lambda^2 & \lambda^3 \\ 0 & \lambda\nu & \lambda^2\nu & \lambda^3\nu \\ 0 & 0 & \lambda^2\nu^2 & \lambda^3\nu^2 \\ 0 & 0 & 0 & \lambda^3\nu^3 \end{pmatrix} \end{array} \begin{matrix} B_0 \\ B_1 \\ B_2 \\ B_3 \end{matrix}$$

where $\nu < 2$ represents the change in adsorption rate caused by the loss of a single sugar from bacterial LPS complex and $\lambda < 1$ is the corresponding change of adsorption rate caused by alterations in the structure of phage tail-fibre protein (Supplementary Information 2.3.1).

We also incorporate two well-established trade-offs: increasing the range of resistance to phages leads to a decrease in growth rate[14–16] and increasing the number of hosts that a phage can infect comes at a reproductive cost through a combination of trade-offs with adsorption rate and burst size[14,17].

To predict bacterial densities we then use equation (2):

$$\begin{aligned} \frac{dS}{dt} &= D(S_0 - S) - c\boldsymbol{\mu}(S)\mathbf{B}^T \\ \frac{d\mathbf{B}}{dt} &= M_b(\boldsymbol{\mu}(S)\cdot\mathbf{B}) - (\Phi\mathbf{P})\cdot\mathbf{B} - D\mathbf{B} \\ \frac{d\mathbf{P}}{dt} &= M_p(\boldsymbol{\beta}\cdot(\Phi^T\mathbf{B})\cdot\mathbf{P}) - D\mathbf{P} \end{aligned} \quad (2)$$

where $\mathbf{B}$ denotes the vector of four bacterial densities and $\mathbf{P}$ denotes the vector of four phage densities. The first equation of (2) describes the rate of change of resource concentration in the chemostat S, with D representing the dilution rate and S_0 representing resource concentration in the input vessel. The consumption of resources is modelled through Michaelis–Menten bacterial growth function $\boldsymbol{\mu}$ and resource conversion rate c whereas phage production is represented by a vector of burst sizes $\boldsymbol{\beta}$ (latent period was not explicitly modelled). The information regarding bacterial and phage mutations is embedded within 4×4 matrices M_b and M_p, respectively, whereas Φ^T represents the transpose of the adsorption matrix Φ (for further description of the model, see Supplementary Information 2).

We fine-tuned the model by using experimentally observed mean rank ordering of bacterial types, obtained as follows. We co-evolved populations of *E. coli* and phages in chemostats and then evaluated the phenotypic diversity of the phage-resistant hosts and phage density. We screened T7-resistant hosts as B_1, B_2 and B_3 on the basis of resistance or sensitivity to a series of reference bacteriophages (Table 1). From the experimental data, we estimated ν to be 0.636 and λ to be 0.94, within the bounds of prior expectations.

The model predicts important differences between different environments (Fig. 2c). First, with resource input around 10 μg ml^{-1} of glucose, the experimental results should be more variable than under high resource input. Thus, we predict that the same bacteria should be found at more similar frequencies in high resource replicates; this is seen (all data, Wilcoxon test, $P = 0.0015$; day 17 data, $P = 0.03$; Fig. 3a). Similarly, the model correctly predicts a higher degree of variation between resource levels than between replicates with the same resource level (Wilcoxon test, $P = 0.003$).

The model also predicts higher phage densities in the higher resource input experiment (Fig. 4), which is observed (Wilcoxon test, $P = 0.0003$; controlling for bacterial type, $P = 0.02$; Fig. 3b). At low resource input, phage type P_2 is predicted to be in greatest abundance, followed by P_3, with other types present at much lower levels (Fig. 4c). Given that both P_2 and P_3 have highest adsorption rate on B_1, the model predicts type B_1 would have higher rates of infection than its competitors. This prediction qualitatively agrees with our experimental findings although, given the rarity of B_1, we could not establish significance (Fig. 3b). In contrast, at high resource input the phage type P_3 is predicted to be most abundant (Fig. 4c). Because this type has higher adsorption rates on B_2 than B_3, the model predicts that B_2 bacteria would be most infected, again in agreement with observation (Wilcoxon test, $P = 0.008$; Fig. 3b).

Given the concordance between theory and observation described above, we conclude that our deterministic mathematical model is fit for purpose. Next we consider the expected outcome if the matrix Φ is

Table 1 | Bacterial phenotypic diversity

Bacteria	Phage			
	T7 WT	T4	T2	Tu1a
B_0 wild type	S	S	S	S
B_1	R	S	S	R
B_2	R	R	S	S
B_3	R	R	R	R

Shown is the bacteriophage screen used to designate host phenotypes. R, resistant; S, sensitive. The bacteriophage screen determines where mutations probably occurred in the T7-resistant bacteria. Bacteriophage T4, mutation LPS; bacteriophage T2, mutation ompF or LPS; bacteriophage Tu1a, mutation ompF.

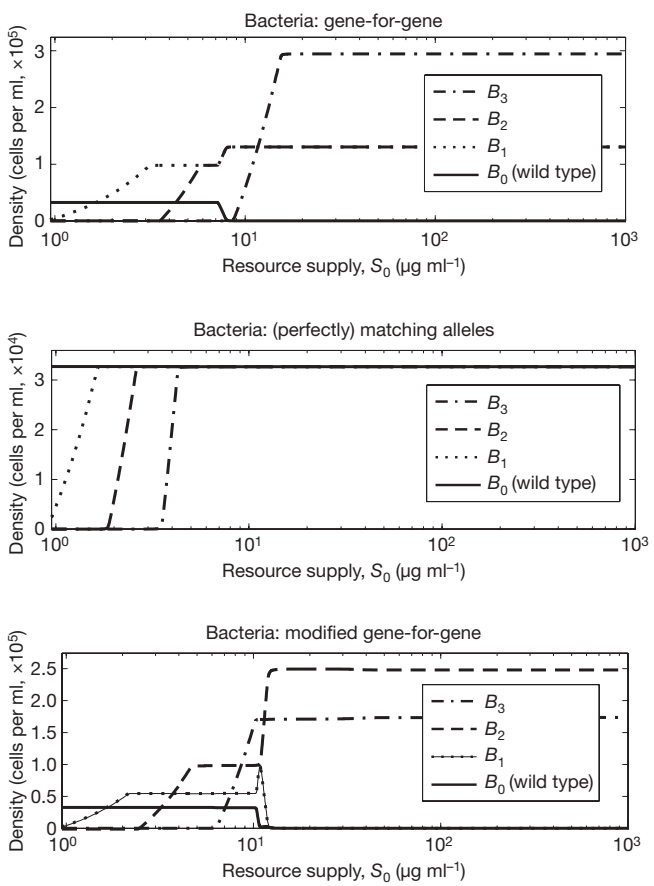


Figure 2 | Bacterial diversity at steady state as a function of resource input, as provided by the mathematical model for different infection mechanisms. **a**, A gene-for-gene model with costs of infection and virulence (infection matrix Φ motivated by ref. 6, with parameter $k = 1/2$ and burst sizes $\beta_0 = 304$, $\beta_1 = 153$, $\beta_2 = 153$, $\beta_3 = 72$). **b**, Matching alleles[6], as found in lambda–*E. coli* using four equal burst sizes of 304. **c**, Modified gene-for-gene model. All other parameter values are given in Supplementary Table 1; bacterial densities are denoted B_i, with i taking values from 0 to 3 and B_0 denoting wild type. We computed these curves taking S_0 from the minimal value required to support phages up to 1,000 μg ml^{-1}.

specific to alternative host–parasite interactions anywhere along the continuum from gene-for-gene to matching alleles[5,7,14,18–20]. Importantly, in all instances of matching allele models, the system is predicted to behave differently from the modified gene-for-gene model presented here, with high diversity observed at high resource input (for discussion see Supplementary Information 6.1; a specific example is shown in Fig. 2b and Supplementary Fig. 8). The interaction between lambda and *E. coli*, in contrast to T7, is a form of lock and key mechanism[18] with phages evolving progressively towards increased affinity for the host receptor[19], hence this is an instance of a matching-allele-like interaction (Fig. 1b). Permitting weak infectivity to non-matching types does not alter the conclusion that diversity should be high at high resource input (see Supplementary Information 6.1). Certain gene-for-gene type matrices, a class considered common in many host–parasite interactions[6,14], can also give this result (Fig. 2a). However there are other matrices that might be deemed gene-for-gene that give the alternative result (see Supplementary Information 6.2).

More generally, we can show that as resource input increases from low to high, we observe two classes of outcome: a monotonic form as predicted for matching-allele-like interactions (for example, with lambda phages) in which diversity of bacteria is highest under high resource input, and the inverted U-shaped form exemplified by T7 in which diversity is maximal at intermediate resource input. The extent to which diversity is below the maximum is dependent on the precise form of Φ and $\boldsymbol{\beta}$. All models predict that the diversity of co-evolving hosts and parasites should vary with differences in resource input (for example, Fig. 2a–c). This contrasts with evolution in the absence of phages, where there is always no change in bacterial diversity with resource input. Because different environments probably provide different resource inputs, creating a selection mosaic, co-evolution of phages and bacteria could drive between-environment differences in diversity, as conceived by the geographic mosaic theory of co-evolution[2]. The passage to a position of stasis at high resource input and the low diversity seen in the T7 case are not particular to assumptions about the number of alleles. So as long as the matrix Φ is square and invertible, the number of alleles has no effect on the available spectrum of diversity properties resulting from the mathematical model (see Supplementary Information 3).

Why is diversity in our study maximal at intermediate resource input? For phages to persist after lysis they need to be able to re-infect bacteria. At very low nutrient levels ($S_0 < 1$) bacterial density is so low that re-infection is unlikely, so phages cannot persist. Consequently the fastest growing bacteria alone are found. As resource supply increases ($10 > S_0 > 1$) bacterial density increases, leading to a zone where intrinsic bacterial growth advantage favours B_0 and B_1, whereas greater resistance to phages favours B_2 and B_3. With phages not dominating the system, this balance potentially enables all four bacterial types to be maintained and for the outcome to be sensitive to S_0. For $S_0 > 10$, bacterial growth and density are both high, but B_0 and B_1 are both killed by phages. With bacteria acting essentially as

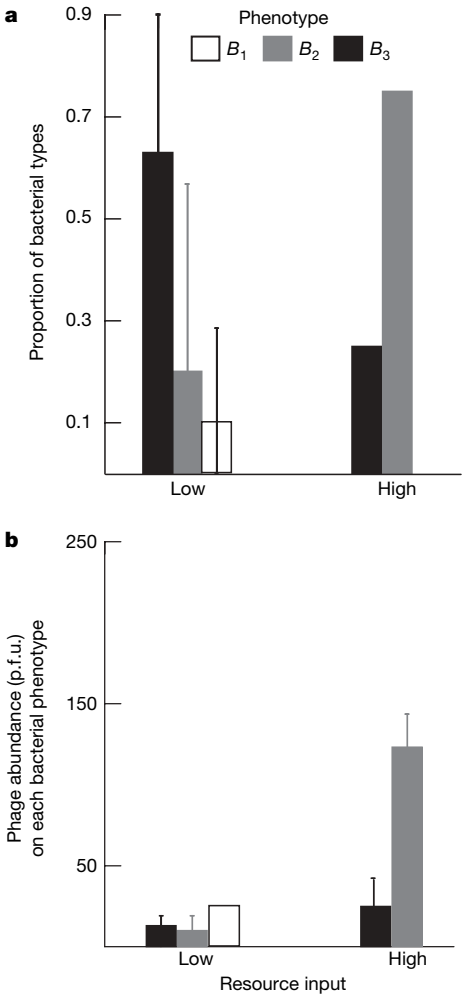


Figure 3 | Experimentally derived bacterial diversity and phage abundance as a function of resource input. **a**, Diversity of bacterial phenotypes. Error bars, s.e.m. **b**, The abundance of phages on each bacterial type (± s.e.m.) from day 17 of the experiment. p.f.u., plaque-forming units.

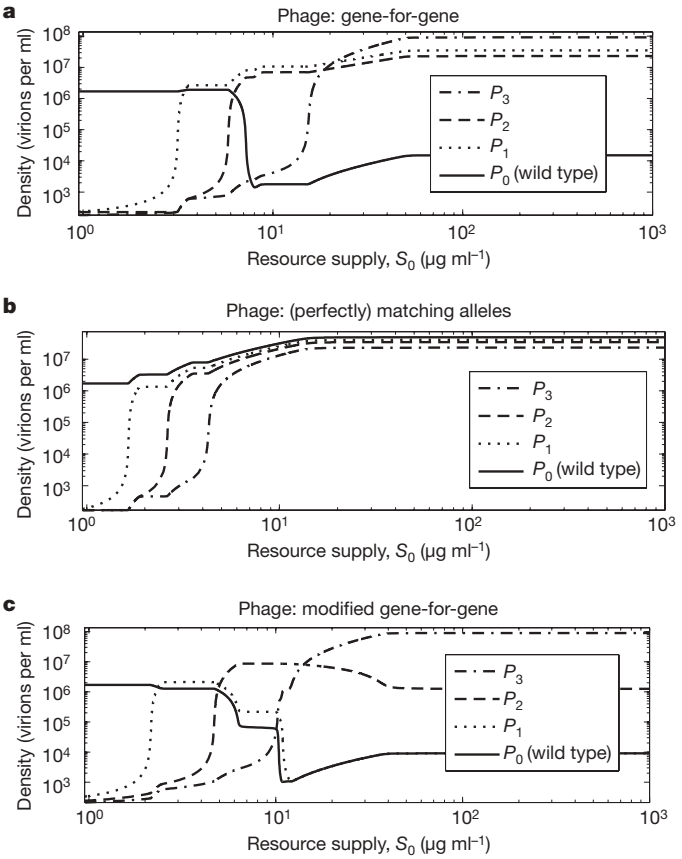


Figure 4 | Phage diversity at steady state as a function of resource input for different infection mechanisms. **a**, A gene-for-gene model with costs of infection and virulence (infection matrix Φ motivated by ref. 6, with parameter $k = 1/2$ and burst sizes $\beta_0 = 304$, $\beta_1 = 153$, $\beta_2 = 153$, $\beta_3 = 72$). **b**, Matching alleles[6], as found in lambda–*E. coli* using four equal burst sizes of 304. **c**, Modified gene-for-gene model. Phage densities are denoted by P_i, where i takes values from 0 to 3 and P_3 denotes the phage type that can infect all bacterial types. We computed these curves taking S_0 from the minimal value required to support phages up to 1,000 μg ml^{-1} (see Supplementary Table 1 for parameter values).

machines converting glucose to phages, these two relatively sensitive types can no longer persist. The resultant diversity is then a balance between differences in growth rate and differences in phage resistance of B_2 and B_3. Increasing bacterial growth rates are kept in check by increasing death rates, so bacterial density changes little. In contrast, in some alternative models, such as matching alleles, we do not necessarily see the removal of B_0 and B_1 because the ancestral types are not sensitive, hence diversity remains high. Given the above explanation, it is perhaps not surprising in retrospect that what is found for T7–*E. coli* interactions need not be true for other biologically viable modes of host–parasite co-evolution. These results show how appropriately framed mathematical models aligned with experimental analysis can obviate the need to presume typicality of one model within a class.

METHODS SUMMARY

The models. The model is based on systems of ordinary differential equations that generate a dissipative dynamical system with a genotype-by-genotype-by-environment structure[2]. The model was parameterized with data on *E. coli* and T7 (Supplementary Information 5). Long-term diversity was computed using a standard Newton-continuation algorithm and applied to the model in steady-state form.

The experiment. Thirty-millilitre communities were inoculated with isogenic strains of *E. coli* and of T7 in chemostats. High resource (1,000 μg ml^{-1} glucose) and low resource (10 μg ml^{-1} glucose) communities were established. Samples of the phage populations and T7-resistant hosts were isolated after 150 bacterial generations of the experiment. T7-resistant colonies were isolated by taking 10 μl from each community, plating it with 50 μl of the ancestral strain of T7 on agar plates and incubating the combined sample at 37 °C overnight (24 h). Each colony was streaked on an agar plate to remove any residual T7 present in the cells and then grown overnight in the same medium as used in the pertinent community. T7-resistant colonies were screened using a series of phages that target the LPS and specific OMPs (Table 1). The bacteriophage screen determined whether changes in the resistance cells had affected LPSs and/or OMPs. We determined the abundance of phages on different bacterial types present in the communities by adding 30 μl of chloroform to 1,000 μl of a sample taken from each community and vortexing the mixture to kill any bacteria that were present. One hundred microlitres of each sample of the phage population was plated on a lawn of each bacterial isolate to determine the abundance of phages on each of the three host phenotypes. Phages were plated on bacterial isolates from the same chemostat from which they originated. The 'efficiency of plating' was our measure of phage abundance.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 31 January; accepted 5 June 2008.

1. Jessup, C. M. *et al.* Big questions, small worlds: microbial model systems in ecology. *Trends Ecol. Evol.* **19,** 189–197 (2004).
2. Thompson, J. *The Geographic Mosaic of Coevolution* (Chicago Univ. Press, Chicago, 2005).
3. Kruger, D. H. & Schroeder, C. Bacteriophage *T3* and bacteriophage *T7* virus–host cell interactions. *Microbiol. Rev.* **45,** 9–51 (1981).
4. Wade, M. J. The co-evolutionary genetics of ecological communities. *Nature Rev. Genet.* **8,** 185–195 (2007).
5. Buckling, A. & Rainey, P. B. Antagonistic coevolution between a bacterium and a bacteriophage. *Proc. R. Soc. Lond. B* **269,** 931–936 (2002).
6. Agrawal, A. & Lively, C. M. Infection genetics: gene-for-gene versus matching-alleles models and all points in between. *Evol. Ecol. Res.* **4,** 79–90 (2002).
7. Morgan, A. D., Gandon, S. & Buckling, A. The effect of migration on local adaptation in a coevolving host–parasite system. *Nature* **437,** 253–256 (2005).
8. Chao, L., Levin, B. R. & Stewart, F. M. Complex community in a simple habitat — experimental-study with bacteria and phage. *Ecology* **58,** 369–378 (1977).
9. Forde, S. E., Thompson, J. N. & Bohannan, B. J. M. Adaptation varies through space and time in a coevolving host–parasitoid interaction. *Nature* **431,** 841–844 (2004).
10. Qimron, U., Marintcheva, B., Tabor, S. & Richardson, C. C. Genomewide screens for *Escherichia coli* genes affecting growth of *T7* bacteriophage. *Proc. Natl Acad. Sci. USA* **103,** 19039–19044 (2006).
11. Sasaki, A. & Godfray, H. C. J. A model for the coevolution of resistance and virulence in coupled host–parasitoid interactions. *Proc. R. Soc. Lond. B* **266,** 455–463 (1999).
12. Tamaki, S., Sato, T. & Matsuhas, M. Role of lipopolysaccharides in antibiotic resistance and bacteriophage adsorption of *Escherichia coli K-12. J. Bact.* **105,** 968–975 (1971).
13. Sen, K. & Nikaido, H. Lipopolysaccharide structure required for *in vitro* trimerization of *Escherichia coli Ompf* porin. *J. Bact.* **173,** 926–928 (1991).
14. Poullain, V., Gandon, S., Brockhurst, M. A., Buckling, A. & Hochberg, M. E. The evolution of specificity in evolving and coevolving antagonistic interactions between a bacteria and its phage. *Evolution* **62,** 1–11 (2008).
15. Bohannan, B. J. M., Kerr, B., Jessup, C. M., Hughes, J. B. & Sandvik, G. Trade-offs and coexistence in microbial microcosms. *Anton Leeuw. Int. J. G.* **81,** 107–115 (2002).
16. Yoshida, T., Hairston, N. G. & Ellner, S. P. Evolutionary trade-off between defence against grazing and competitive ability in a simple unicellular alga, *Chlorelia vulgaris. Proc. R. Soc. Lond. B* **271,** 1947–1953 (2004).
17. Ferris, M. T., Joyce, P. & Burch, C. L. High frequency of mutations that expand the host range of an RNA virus. *Genetics* **176,** 1013–1022 (2007).
18. Weitz, J. S., Hartman, H. & Levin, S. A. Coevolutionary arms races between bacteria and bacteriophage. *Proc. Natl Acad. Sci. USA* **102,** 9535–9540 (2005).
19. Spanakis, E. & Horne, M. T. Co-adaptation of *Escherichia coli* and coliphage *lambda vir* in continuous culture. *J. Gen. Microbiol.* **133,** 353–360 (1987).
20. Lenski, R. E. & Levin, B. R. Constraints on the coevolution of bacteria and virulent phage — a model, some experiments, and predictions for natural communities. *Am. Nat.* **125,** 585–602 (1985).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Buckling, S. Nuismer, K. Rich, J. Hoeksema and C. Jessup for their comments on an earlier version of this manuscript. L.D.H. is a Royal Society Wolfson Research Merit Award Holder. I.G. is supported by a NERC Advanced Fellowship. S.S.A. is funded by an ORS award and a studentship for the Department of Mathematics at Imperial College London. S.E.F. and J.N.T. are supported by the National Science Foundation DEB 0515598.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to L.D.H. (l.d.hurst@bath.ac.uk)

METHODS

The models. The model is based on systems of ordinary differential equations that each generate a dissipative dynamical system with a genotype-by-genotype-by-environment structure[2]. The model was parameterized with data on *E. coli* and T7 (Supplementary Information 5), and we concluded that there exists a globally attractive state of equilibrium densities that can be approached in an oscillatory manner. Thus, one can summarize long-term diversity by plotting resource input S_0 versus the equilibrium densities of bacterial and phage types computed using a standard Newton-continuation algorithm implemented in Matlab and applied to the model in steady-state form.

The experiment. Thirty-millilitre communities were inoculated with isogenic strains of *E. coli* and of T7 in chemostats. Two types of communities were established by manipulating the input of limiting nutrients for the bacteria: high resource ($1{,}000\,\mu g\,ml^{-1}$ glucose; three communities) and low resource ($10\,\mu g\,ml^{-1}$glucose; two communities). Samples of the phage populations and T7-resistant hosts were isolated after the initial invasion of the resistant mutants and after the host and parasitoid co-evolved for more than 150 bacterial generations of the experiment (initial sample, high resource = 19, low resource = 11; final sample, high = 11, low = 12 bacterial colonies across all chemostats). See ref. 21 for average population sizes.

Phenotypic diversity of resistant hosts. T7-resistant colonies were isolated by taking 10 μl from each community, plating it with 50 μl of the ancestral strain of T7 (titre of approximately 1×10^8) on agar plates, and incubating the combined sample at 37 °C overnight. Note that measuring the phenotypic diversity of the resistant hosts guarantees that selection has occurred, and thus any phage that can attack the hosts must be host-range mutants. Each colony was then streaked on an agar plate to remove any residual T7 present in the cells and grown overnight in the same type of liquid medium as that used in the original experiment (that is, either high or low resources). Freezer stocks of each culture were then stored in glycerol at −80 °C for future use.

T7-resistant colonies were then screened using a series of phages that target the LPS and specific OMPs (Table 1). The presence of LPSs is involved in maintenance of cell integrity and impermeability whereas OMPs are involved in uptake of nutrients into the cell and in outer membrane stability. The subscripts of the four types (B_0–B_3) refer to the number of LPS-targeting reference phages to which the bacterial type is resistant (see Table 1) and also orders the types according to their growth kinetics, with B_0 having the highest and B_3 the lowest growth rate (see Supplementary Information 5).

Each bacterial isolate was grown overnight in the appropriate medium (high or low resource) and then streaked across 20 μl of each reference phage that had been dried on an agar plate to assess resistance. In combination, these screens allowed us to determine bacterial phenotypes in the high and low resource communities. The proportions of each phenotype at the start and end of the experiment were then averaged over time.

Abundance of phages on different bacterial phenotypes. We determined the abundance of phages on different bacterial types present in the communities by adding 30 μl of chloroform to 1,000 μl of a sample taken from each community and vortexing the mixture to kill any bacteria that were present. One hundred microlitres of each sample of the phage population was plated on a lawn of each bacterial isolate to determine the abundance of phages on each of the three host phenotypes (B_1, B_2 and B_3). Phages were also plated on ancestral bacteria (B_0). The number of phage plaques was consistently higher on B_0. Phages were plated on bacterial isolates from the same chemostat from which they originated (5–7 isolates per chemostat), and we used the 'efficiency of plating' (the number of plaques on each host) as a measure of phage abundance.

Data analysis. We examined the prediction that there should be higher repeatability in experimental outcome at high resource input, for each of three bacterial types (B_1, B_2, B_3) by considering the modulus of the difference in the frequency of each type in each replicate experiment in a given resource level. In all replicates at high resource input there is no difference in the frequency of each bacterial type. At low input the mean difference in frequency between replicate experiments is 0.21 ± 0.08 (s.e.m.), which is significantly greater than seen at high resource input (all data: Wilcoxon test $P = 0.0015$; day 17 data: $P = 0.03$).

We asked whether there is more variation between levels than within by considering the modular difference in frequency of the same bacterial type between resource input levels, and asked whether it is greater than the differences observed within resource levels, as predicted by the model: the within resource level mean modular difference is 0.1 ± 0.03 (s.e.m.), which is lower than the mean modular differences of frequencies observed at day 17 between the same bacterial types at difference resource levels (mean modular difference in bacterial frequency between resource levels is 0.37 ± 0.13: Wilcoxon test $P = 0.003$).

21. Forde, S. E., Thompson, J. N. & Bohannan, B. J. Gene flow reverses an adaptive cline in a coevolving host-parasitoid interaction. *Am. Nat.* **169**, 794–801 (2007).

LETTERS

High bacterivory by the smallest phytoplankton in the North Atlantic Ocean

Mikhail V. Zubkov[1] & Glen A. Tarran[2]

Planktonic algae <5 μm in size are major fixers of inorganic carbon in the ocean[1]. They dominate phytoplankton biomass in post-bloom, stratified oceanic temperate waters[2]. Traditionally, large and small algae are viewed as having a critical growth dependence on inorganic nutrients, which the latter can better acquire at lower ambient concentrations owing to their higher surface area to volume ratios[3,4]. Nonetheless, recent phosphate tracer experiments in the oligotrophic ocean[5] have suggested that small algae obtain inorganic phosphate indirectly, possibly through feeding on bacterioplankton. There have been numerous microscopy-based studies of algae feeding mixotrophically[6,7] in the laboratory[8–10] and field[11–14], as well as mathematical modelling of the ecological importance of mixotrophy[15]. However, because of methodological limitations[16] there has not been a direct comparison of obligate heterotrophic and mixotrophic bacterivory. Here we present direct evidence that small algae carry out 40–95% of the bacterivory in the euphotic layer of the temperate North Atlantic Ocean in summer. A similar range of 37–70% was determined in the surface waters of the tropical North-East Atlantic Ocean, suggesting the global significance of mixotrophy. This finding reveals that even the smallest algae have less dependence on dissolved inorganic nutrients than previously thought, obtaining a quarter of their biomass from bacterivory. This has important implications for how we perceive nutrient acquisition and limitation of carbon-fixing protists as well as control of bacterioplankton in the ocean.

We carried out the study on board the Royal Research Ship *Discovery* in the central temperate North Atlantic Ocean at 58.2–60° N and 18.7–22° W (Supplementary Fig. 1) between 30 July and 19 August 2007. Bacterivory by planktonic plastidic, phototrophic protists (algae) and aplastidic, heterotrophic protists <5 μm in size was compared in 13 experiments using samples collected from the depths of either 7 m (surface mixed layer, 9 stations) or 47 m (thermocline, 4 stations). A supplementary study of bacterivory (at 20 m, 3 stations) was performed on board the same ship in the tropical North-East Atlantic Ocean at 12.6–22° N and 27–33° W (Supplementary Fig. 2) on 21–30 January 2008.

The experimental approach, developed on laboratory cultures[17,18], allowed the determination of the proportion of bacterioplankton biomass (pulse-chase labelled with ^{35}S-methionine and ^{3}H-leucine), which was assimilated by protists. The chase, with thousand-fold higher concentrations of non-labelled tracer analogues compared with the tracer pulses, effectively stopped direct uptake of radiolabelled molecules by all microbial cells. To quantify indirect acquisition of radiotracers by the dominant protist groups through bacterivory, flow cytometric sorting of pulse-chase labelled microbes was used. Cells representing the major microbial planktonic groups (Supplementary Fig. 3)—aplastidic protists (~5 μm) as well as large (~5 μm), medium (~3 μm) and small (~2 μm) plastidic protists—and bacterioplankton were flow cytometrically sorted.

The cells were sorted from samples collected at two to four time points. From statistically significant differences (*t*-test, $P < 0.05$) in cellular radioactivity at these times (Supplementary Figs 6 and 7), we were able to calculate bacterioplankton biomass assimilation by protist cells. To compare assimilation rates at different stations, the time differences in protist cell radioactivity were divided by the corresponding mean radioactivity of an average bacterioplankton cell from the same grazing experiment and presented as bacterioplankton cell equivalents assimilated by a protist cell per hour (Fig. 1a). Population-specific bacterioplankton assimilation of flow-sorted protist groups was determined by multiplying assimilation of an average cell by the abundance of the group (Fig. 1b, c).

Although during the chase period the retention of ^{3}H-leucine and ^{35}S-methionine tracers by bacterioplankton cells was similarly good (Supplementary Figs 5–7), the assimilation rates of bacterioplankton biomass by protist cells, determined using the ^{35}S tracer, were consistently higher than the assimilation rates determined using the ^{3}H tracer: from 2.5 times for aplastidic protists to 4 times for large and small plastidic protists. Presumably, the ^{3}H-multiple-labelled part of the leucine molecule was more readily metabolized by protists during prey digestion than ^{35}S-labelled methionine. As a result, we used the statistically more robust ^{35}S tracer data to calculate and to compare assimilation rates of bacterioplankton biomass by different groups of protists. Because some of the ^{35}S-labelled molecules, taken up by protist cells, could also be metabolized and the ^{35}S tracer released by cells, the calculated assimilation rates provided conservative estimates of the rates of protist bacterivory.

Aplastidic protists assimilated bacterioplankton biomass at a rate equivalent to 1.5–5.0 average bacterioplankton cells per protist cell per hour (Fig. 1a). Being specialized bacterivores, aplastidic protists had the highest bacterivory rates among planktonic protists both in the surface mixed layer and the thermocline: on average it was 13 times higher ($P < 0.001$) than small plastidic protists, 8.3 times higher ($P < 0.005$) than medium plastidic protists and 3.9 times higher ($P < 0.001$) than large plastidic protists in the temperate waters. Considering that cell sizes of aplastidic and large plastidic protists were similar, they probably required similar cell biomass to be synthesized for a cell division. Aplastidic protists should derive this biomass entirely from bacterivory. Hence, the large plastidic protists obtain about 25% of their biomass from bacterivory, given the above rates. Also, in the temperate waters, large plastidic protists showed 2.3 times higher bacterivory rates in the thermocline ($P < 0.03$), becoming more bacterivory-dependent at depth when compared to the surface waters. The bacterivory rates of aplastidic protists in the surface mixed layer and in the thermocline were statistically similar, and the same was true of small plastidic protists.

Although the cellular bacterivory rates of plastidic protists were lower than the cellular bacterivory rates of aplastidic protists (Fig. 1a), the former were more abundant (Fig. 1b) so that the cumulative

[1]National Oceanography Centre, Southampton, Hampshire SO14 3ZH, UK. [2]Plymouth Marine Laboratory, Plymouth, Devon PL1 3DH, UK.

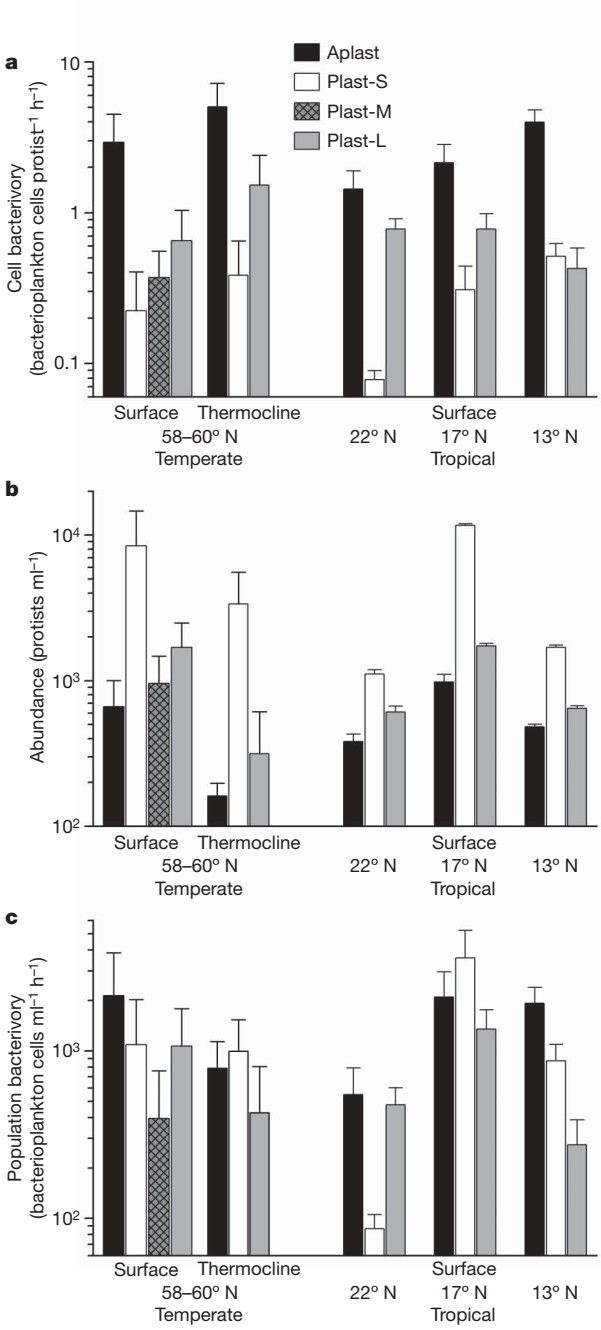


Figure 1 | Bacterivory rates and abundance of protist groups. Comparison of bacterivory by protist groups in the surface mixed layer and at the thermocline in the temperate waters and in the surface mixed layer in the tropical waters of the North Atlantic Ocean (latitude, degrees N) based on ^{35}S-methionine pulse-chase tracing and bacterioplankton cell equivalents. The groups are: aplastidic protists (Aplast) and large, medium and small plastidic protists (Plast-L, Plast-M and Plast-S, respectively). **a**, Cellular bacterivory rates. **b**, Abundance of protist groups. **c**, Population bacterivory rates. Error bars indicate single standard deviations of average values of nine (surface) or four (thermocline) repeated experiments, carried out in the temperate waters. A single experiment was carried out at each of the three latitudes in the tropical waters; the error bars indicate single standard deviations of average values of 3–4 replicated measurements of either protist bacterivory or protist abundance, that is, experimental errors of measurements. Determination of bacterivory of the Plast-M group in the thermocline in the temperate waters was not practical because of the low abundance of the group combined with the low cellular rates of tracer assimilation. The Plast-M group was not pronounced in the tropical waters.

bacterivory rates of plastidic protist populations were comparable to the bacterivory rates of the aplastidic protist population (Fig. 1c). The summed bacterivory rates of plastidic protist populations were either equal to, or higher than, the bacterivory rates of aplastidic protist populations. Thus, populations of plastidic protists, which dominated phytoplankton in the temperate waters (the $<5\,\mu m$ fraction accounted for $>80\%$ of total phytoplankton chlorophyll, M. Moore, personal communication), were responsible for 40–95% of total bacterivory; on average it was statistically similar in the surface mixed layer and in the thermocline, representing $51 \pm 11\%$ (mean $\pm$ standard deviation of nine experiments) and $61 \pm 16\%$ of bacterivory (mean $\pm$ standard deviation of four experiments). Similarly, populations of plastidic protists were responsible for $53 \pm 17\%$ (mean $\pm$ standard deviation of three experiments; 37–70% range) of total bacterivory in the tropical surface waters.

Although taxonomic identification of the dominant protists $<5\,\mu m$ that comprised the sorted groups is beyond the scope of this communication, we expect mixotrophic haptophytes to be abundant. However, we do not exclude that other common protists—for example, small prasinophytes, which grow as obligate phototrophs in nutrient-rich, laboratory cultures—may be mixotrophs in nutrient-depleted, oceanic waters and feed on 'unculturable' bacterioplankton. Making a parallel with studies of marine cyanobacteria, until significant uptake of amino acids was demonstrated in the open ocean[19] *Prochlorococcus* were also considered to be obligate phototrophs. There could be similar trophic surprises among protists.

From a broader perspective, bacterivory can provide plastidic cells with concentrated nutrients, which are readily available for growth on digestion. Assuming that the nitrogen (N) content of an average bacterioplankton cell is 3.5 fg[20], bacterioplankton ($1.7 \pm 0.5 \times 10^6$ cells ml^{-1}; mean $\pm$ standard deviation of 13 measurements) formed a reliable nitrogen pool of approximately 0.3 μmol N l^{-1}, which was comparable to a micromolar pool of inorganic nitrogen in the stratified surface layer of the temperate North Atlantic in summer. By tapping into the bacterioplankton nitrogen pool, the mixotrophic cells acquire a competitive edge over obligate phototrophic cells even when the inorganic nutrients are not strongly depleted, for example in the temperate waters or in the vicinity of the Cape Verde Islands.

In the euphotic layer, mixotrophs could also efficiently compete with heterotrophs, as has been observed in fresh water[13]. In the temperate North Atlantic the high abundance of bacterioplankton sustained a sizeable population of aplastidic protists ($0.51 \pm 0.37 \times 10^3$ cells ml^{-1}; mean $\pm$ standard deviation of 13 measurements), which showed more efficient cell predation than plastidic protists (Fig. 1a). Although the average abundance of aplastidic protists was only 6% of the sum of the abundance of plastidic protists, the former were responsible for up to 60% of the bacterivory, supporting the view that specialization pays dividends. However, if the ecological success of a population is measured not by its efficiency but by its abundance (Fig. 1b), then mixotrophic plastidic protists control planktonic microbial communities of surface waters in the temperate open ocean and probably in the tropical waters as well (Fig. 1c).

Despite being the main control of bacterioplankton abundance, the combined bacterivory by both plastidic and aplastidic plankton accounted for only $0.25 \pm 0.15\%$ of the bacterioplankton standing stock per hour in the temperate waters and for 0.13–0.8% in the tropical waters. Given these low values, it is not surprising why it has been so challenging to develop a technique for determining *in situ* rates of bacterivory by planktonic protists in the open ocean[21–23]. We have for the first time, to our knowledge, measured comparatively the bacterivory rates of open-ocean phototrophic protists and heterotrophic protists, fed on ambient bacterioplankton, and in doing so we have shown the large contribution of phytoplankton in harvesting bacterioplankton, and have challenged the assumption of total dependence of phytoplankton on inorganic nutrients.

METHODS SUMMARY

Pulse-chase labelling. For each experiment a sample of 240 ml sea water was accurately placed into a 250 ml glass bottle cleaned with 10% hydrochloric acid. In all experiments L-[^{35}S]methionine (specific activity 37 TBq mM^{-1}, GE Healthcare) was added at 0.25 nM final concentration. In the initial four experiments in the temperate waters, carried out with surface water samples, methionine was the single tracer. In the latter nine dual tracer experiments, L-[4,5-^{3}H]leucine (specific activity 5.96 TBq mM^{-1}) was added at 0.5 nM final concentration. After 1 h incubation in the dark at *in situ* temperature, 0.25 μM and 0.5 μM final concentrations of non-radioactive L-methionine and L-leucine were added to chase pulses of the radioactive amino acids in synthesized microbial proteins by sharply reducing the specific radioactivity of the tracer molecules (Supplementary Fig. 5). The sample was incubated for an additional hour before commencing the measurements of the amount of bacterioplankton biomass, pulse-chase labelled with ^{35}S-methionine and ^{3}H-leucine, which was assimilated by protists between 2 h and 5 h. Subsamples of 120 ml were fixed with 1% (w/v) paraformaldehyde (PFA) final concentration after 2 h and 5 h. On two occasions, pulse-chased microbial dynamics of tracers was monitored during the entire incubation period (Supplementary Fig. 5), with bacterioplankton and protist cells flow-sorted from samples collected at four time points (Supplementary Fig. 6). In the tropical waters, one experiment using L-[^{35}S]methionine (specific activity 43.5 TBq mM^{-1}, PerkinElmer) as the single tracer was carried out at each location with subsamples fixed after 2 h and 8 h.

Flow cytometric sorting. Bacterioplankton cells were flow-sorted from unconcentrated, SYBR Green I DNA-stained samples[24]. Protists (Supplementary Fig. 3) were generally flow-sorted from 0.8-μm concentrated samples[5]. Radioassaying was performed[5,25] using an ultra-low-level liquid scintillation counter (1220 Quantulus, Wallac).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 21 April; accepted 4 July 2008.
Published online 10 August 2008.

1. Li, W. K. W. Primary production of prochlorophytes, cyanobacteria, and eukaryotic ultraphytoplankton — measurements from flow cytometric sorting. *Limnol. Oceanogr.* **39,** 169–175 (1994).
2. Tarran, G. A., Zubkov, M. V., Sleigh, M. A., Burkill, P. H. & Yallop, M. Microbial community structure and standing stocks in the NE Atlantic in June and July of 1996. *Deep-Sea Res. II* **48,** 963–985 (2001).
3. Malone, T. C. in *The Physiological Ecology of Phytoplankton* (ed. Morris, I.) 433–463 (Blackwell Scientific, 1980).
4. Chisholm, S. W. in *Primary Productivity and Biogeochemical Cycles in the Sea* (eds Falkowski, P. G. & Woodhead, A. D.) 213–237 (Plenum, 1992).
5. Zubkov, M. V. *et al.* Microbial control of phosphate in the nutrient-depleted North Atlantic subtropical gyre. *Environ. Microbiol.* **9,** 2079–2089 (2007).
6. Caron, D. A. in *Microbial Ecology of the Oceans* (ed. Kirchman, D.) 495–523 (Wiley & Sons, 2000).
7. Jones, R. I. Mixotrophy in planktonic protists: an overview. *Freshwater Biol.* **45,** 219–226 (2000).
8. Rothhaupt, K. O. Laboratory experiments with a mixotrophic chrysophyte and obligately phagotrophic and phototrophic competitors. *Ecology* **77,** 716–724 (1996).
9. Stibor, H. & Sommer, U. Mixotrophy of a photosynthetic flagellate viewed from an optimal foraging perspective. *Protist* **154,** 91–98 (2003).
10. Hansen, P. J. & Hjorth, M. Growth and grazing responses of *Chrysochromulina ericina* (Prymnesiophyceae): the role of irradiance, prey concentration and pH. *Mar. Biol.* **141,** 975–983 (2002).
11. Bird, D. F. & Kalff, J. Bacterial grazing by planktonic lake algae. *Science* **231,** 493–495 (1986).
12. Arenovski, A. L., Lim, E. L. & Caron, D. A. Mixotrophic nanoplankton in oligotrophic surface waters of the Sargasso Sea may employ phagotrophy to obtain major nutrients. *J. Plankton Res.* **17,** 801–820 (1995).
13. Tittel, J. *et al.* Mixotrophs combine resource use to outcompete specialists: implications for aquatic food webs. *Proc. Natl Acad. Sci. USA* **100,** 12776–12781 (2003).
14. Unrein, F., Massana, R., Alonso-Saez, L. & Gasol, J. M. Significant year-round effect of small mixotrophic flagellates on bacterioplankton in an oligotrophic coastal system. *Limnol. Oceanogr.* **52,** 456–469 (2007).
15. Thingstad, T. F., Havskum, H., Garde, K. & Riemann, B. On the strategy of "eating your competitor": A mathematical analysis of algal mixotrophy. *Ecology* **77,** 2108–2118 (1996).
16. Vaque, D., Gasol, J. M. & Marrase, C. Grazing rates on bacteria — the significance of methodology and ecological factors. *Mar. Ecol. Prog. Ser.* **109,** 263–274 (1994).
17. Zubkov, M. V. & Sleigh, M. A. Ingestion and assimilation by marine protists fed on bacteria labeled with radioactive thymidine and leucine estimated without separating predator and prey. *Microb. Ecol.* **30,** 157–170 (1995).
18. Zubkov, M. V. & Sleigh, M. A. Assimilation efficiency of *Vibrio* bacterial protein biomass by the flagellate *Pteridomonas*: Assessment using flow cytometric sorting. *FEMS Microbiol. Ecol.* **54,** 281–286 (2005).
19. Zubkov, M. V., Tarran, G. A. & Fuchs, B. M. Depth related amino acid uptake by *Prochlorococcus* cyanobacteria in the Southern Atlantic tropical gyre. *FEMS Microbiol. Ecol.* **50,** 153–161 (2004).
20. Fagerbakke, K. M., Heldal, M. & Norland, S. Content of carbon, nitrogen, oxygen, sulfur and phosphorus in native aquatic and cultured bacteria. *Aquat. Microb. Ecol.* **10,** 15–27 (1996).
21. Sherr, B. F., Sherr, E. B. & Fallon, R. D. Use of monodispersed, fluorescently labeled bacteria to estimate in situ protozoan bacterivory. *Appl. Environ. Microbiol.* **53,** 958–965 (1987).
22. Zubkov, M. V., Sleigh, M. A. & Burkill, P. H. Measurement of bacterivory by protists in open ocean waters. *FEMS Microbiol. Ecol.* **27,** 85–102 (1998).
23. Sherr, E. B. & Sherr, B. F. Significance of predation by protists in aquatic microbial food webs. *Antonie Van Leeuwenhoek* **81,** 293–308 (2002).
24. Marie, D., Partensky, F., Jacquet, S. & Vaulot, D. Enumeration and cell cycle analysis of natural populations of marine picoplankton by flow cytometry using the nucleic acid stain SYBR Green I. *Appl. Environ. Microbiol.* **63,** 186–193 (1997).
25. Mary, I. *et al.* Light enhanced amino acid uptake by dominant bacterioplankton groups in surface waters of the Atlantic Ocean. *FEMS Microbiol. Ecol.* **63,** 36–45 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We gratefully acknowledge the captain, officers and crew aboard the RRS *Discovery* for their help during the cruises, M. Moore for sharing chlorophyll concentration measurements, P. Hill for helping with setting up experiments on the second cruise, and P. Warwick for his help with radiotracer measurements ashore. We thank M. Sleigh, A. Martin and R. Leakey for their critical comments on the earlier drafts of the paper. We thank H. Ducklow for constructive criticism of the earlier version of this paper. This study was supported by the UK Natural Environment Research Council (NERC) through the Oceans 2025 core programmes of the National Oceanography Centre, Southampton and Plymouth Marine Laboratory. The second cruise was supported by the NERC thematic programme Surface Ocean Low Atmosphere Study (SOLAS).

Author Contributions The experimental approach was designed by M.Z., who carried out the tracer work. Flow cytometric analyses were done by G.T. and M.Z., respectively, on the first and second cruise.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.V.Z. (mvz@noc.soton.ac.uk).

METHODS

Sampling. Seawater samples were collected before dawn with 10 l, trace-iron-free Niskin bottles mounted on a titanium conductivity–temperature–depth profiler.

Cell enumeration. Abundances of bacterioplankton and protists (Supplementary Fig. 3) were determined by flow cytometry (FACSort, Becton Dickinson) in samples, fixed with 1% (w/v) paraformaldehyde (PFA) final concentration and stained with SYBR Green I DNA dye[24,26]. Abundances of plastidic protists were also determined in live, unstained samples[1,2,27]. The counts of unstained and stained plastidic protists correlated with a close to 1:1 relationship (Supplementary Fig. 4)[26]. Multifluorescence 0.5 μm reference beads (fluoresbrite microparticles, Polysciences) were used in all analyses as an internal standard for both fluorescence and flow rates. The absolute concentration of beads in the stock solution was determined using syringe pump flow cytometry[28].

Flow cytometric sorting. Pulse-chase-labelled, PFA-fixed samples were stored at 4 °C for microbial flow cytometric sorting, which was carried out within 10 h. Bacterioplankton cells were flow-sorted from SYBR Green I DNA-stained, unconcentrated samples[5,19,25] using a FACSCalibur flow cytometer (Becton Dickinson). Cells representing the protist groups—aplastidic protists as well as large, medium and small plastidic protists (Supplementary Fig. 3)—were flow-sorted from unconcentrated samples, collected in the tropical waters, and from concentrated samples, collected in the temperate waters. To concentrate protist cells, samples of 110 ml sea water, fixed with 1% PFA, were amended with 0.05% pluronic solution (Sigma) final concentration. Samples were then gently filtered through 0.8-μm pore-size polycarbonate filters, diameter 25 mm (Whatmann), housed in Swinnex filter holder units (Millipore) using a syringe pump KDS-230 (KD Scientific) equipped with 60-ml plastic syringes. The concentrated cells were washed off filters and resuspended in 1.8 ml of the unfiltered, fixed sample before being stained with SYBR Green I and then flow-sorted. The FACSCalibur was set at single-cell sort mode (the highest sorting purity of the instrument) and the target cells were gated (Supplementary Fig. 3) and flow-sorted at a rate of 1–250 particles s^{-1}. Sorted cells were collected onto 0.2-μm pore-size polycarbonate filters and washed twice with deionized water.

Radioactivity retained on filters was provisionally assayed using a liquid scintillation counter Tri-Carb 3100 (PerkinElmer) on board the ship. Accurate radioassaying of filters with single and dual tracers was done using an ultra-low-level liquid scintillation counter 1220 Quantulus (Wallac). The latter counter uses a logarithmic analogue-to-digital converter, which effectively expands the low energy end of the spectrum and facilitates more effective deconvolution of the two spectra. Tritium (^{3}H) and ^{35}S counting efficiencies and spillover of the ^{35}S into the ^{3}H counting window were corrected for using a quench-dependent calibration. Disintegrations per minute (d.p.m.) were calculated to correct for the radioactive decay. At least three proportional numbers of cells were sorted from 0.2×10^3, 0.4×10^3 and 0.6×10^3 cells to 20×10^3, 40×10^3 and 60×10^3 cells, and the mean cellular content of tracers was determined.

The total microbial uptake of tracers was determined in samples used for flow-sorting by filtering triplicate sub-samples of 150 μl, 300 μl and 450 μl onto 0.2-μm pore-size polycarbonate filters, washed twice with deionized water and radioassayed as described above. The total microbial uptake rates of ^{3}H-leucine and ^{35}S-methionine after 2 h pulse-chase were $6.7 \pm 3.7\%$ ($n = 9$) and $4.3 \pm 1.9\%$ ($n = 13$; mean $\pm$ standard deviation) of the added amount of the two tracers, respectively.

Statistically significant differences (t-test, $P < 0.05$) of cellular radioactivity (Supplementary Fig. 7) at the sampling times (2 h and 5 h in temperate waters; 2 h and 8 h in tropical waters) were used for calculation of bacterioplankton biomass assimilation by protist cells. To compare assimilation rates at different stations, the time differences in protist cell radioactivity were divided by corresponding mean radioactivity of an average bacterioplankton cell and presented as bacterioplankton cell equivalents assimilated by a protist cell per hour, considering linearity of protist label acquisition (Supplementary Fig. 6). Population-specific bacterioplankton assimilation of flow-sorted protist groups was determined by multiplying assimilation of their average cell by the abundance of the group.

Monitoring quality of flow-sorting. Before starting to sort radiotracer-labelled cells, the sorter alignment was checked by sorting one type of beads from a mixture of two 0.5 μm beads with different yellow-green fluorescence. The sorted material was 99.8% enriched with the target beads; the sorted bead recovery was $98.8 \pm 0.9\%$ ($n = 7$; mean $\pm$ standard deviation). Two radiotracer budget controls were used to monitor sorting precision throughout the cruise. First, the tracer radioactivity of an average bacterioplankton cell was multiplied by the concentration of bacterioplankton in the sample. The resulting total bacterioplankton population radioactivity was compared with the tracer radioactivity of particulate material, collected from the same sample directly onto a 0.2-μm pore-size filter. The two data sets correlated with a close to 1:1 relationship for both ^{3}H and ^{35}S tracers (Supplementary Fig. 8a, b), validating the high accuracy of flow cytometric counting and sorting. Second, we compared the tracer radioactivities of the total bacterioplankton population and the sum of the two main bacterioplankton subpopulations. The cells with low nucleic acid (LNA) and high nucleic acid (HNA) content were sorted in parallel with average bacterioplankton cells. The tracer radioactivities of the two subpopulations were calculated by multiplying the mean cellular radioactivities of a LNA cell or a HNA cell by the corresponding LNA or HNA cell abundances. The relationship between the radioactivities of the total bacterioplankton population and the sum of the two subpopulations was also very close to 1:1 for both ^{3}H and ^{35}S tracers (Supplementary Fig. 9a, b), corroborating the high precision of flow-sorting.

F-tests and t-tests were used, respectively, for comparison of variance and means of various data sets.

26. Zubkov, M. V., Burkill, P. H. & Topping, J. N. Flow cytometric enumeration of DNA-stained oceanic planktonic protists. *J. Plankton Res.* **29**, 79–86 (2007).
27. Olson, R. J., Zettler, E. R. & DuRand, M. D. in *Handbook of Methods in Aquatic Microbial Ecology* (eds Kemp, P. F., Sherr, B. F., Sherr, E. B. & Cole, J. J.) 175–186 (Lewis, 1993).
28. Zubkov, M. V. & Burkill, P. H. Syringe pumped high speed flow cytometry of oceanic phytoplankton. *Cytometry A* **69A**, 1010–1019 (2006).

LETTERS

Neural correlates, computation and behavioural impact of decision confidence

Adam Kepecs[1], Naoshige Uchida[1,2], Hatim A. Zariwala[1,3] & Zachary F. Mainen[1,4]

Humans and other animals must often make decisions on the basis of imperfect evidence[1,2]. Statisticians use measures such as *P* values to assign degrees of confidence to propositions, but little is known about how the brain computes confidence estimates about decisions. We explored this issue using behavioural analysis and neural recordings in rats in combination with computational modelling. Subjects were trained to perform an odour categorization task that allowed decision confidence to be manipulated by varying the distance of the test stimulus to the category boundary. To understand how confidence could be computed along with the choice itself, using standard models of decision-making[3–6], we defined a simple measure that quantified the quality of the evidence contributing to a particular decision. Here we show that the firing rates of many single neurons in the orbitofrontal cortex match closely to the predictions of confidence models and cannot be readily explained by alternative mechanisms, such as learning stimulus–outcome associations[7–10]. Moreover, when tested using a delayed reward version of the task, we found that rats' willingness to wait for rewards increased with confidence, as predicted by the theoretical model. These results indicate that confidence estimates, previously suggested to require 'metacognition'[11,12] and conscious awareness[13,14], are available even in the rodent brain, can be computed with relatively simple operations, and can drive adaptive behaviour. We suggest that confidence estimation may be a fundamental and ubiquitous component of decision-making.

Rats were trained on a two choice odour mixture categorization task (Fig. 1a). On each trial, a binary mixture of two pure odorants (A, caproic acid; B, 1-hexanol) was delivered at one of several concentration ratios (Fig. 1b), which were randomly interleaved from trial-to-trial[15]. Choices were rewarded at the left choice port for mixtures A/B > 50/50 and at the right choice port for A/B < 50/50 (Fig. 1b). By varying the distance of the stimulus to the category boundary (50/50) we could vary the difficulty of the decision (Fig. 1c, d). Although the reward contingencies were deterministic, subjects experienced varying degrees of decision uncertainty due to imperfect perception of stimuli and/or knowledge of the category boundary.

To explore the neural correlates of decision confidence, we recorded single neuron activity in the orbitofrontal cortex (OFC; Supplementary Fig. 1), a brain region implicated in decision-making under uncertainty[16–20]. We reasoned that neural activity related to the subject's confidence in the outcome of a choice should occur while the subject is anticipating the trial outcome, and therefore focused our analysis on this delay period (Fig. 2a). The firing rates of many OFC neurons were modulated by stimulus difficulty during the anticipation period. Figure 2b, c shows the activity of a neuron that fired more intensely following more difficult decisions. By replotting the same data as a function of the choice accuracy associated with each stimulus type, it can be seen that this neuron fired more vigorously when the likelihood of an upcoming reward was lower (Fig. 2d). A large fraction of OFC neurons, like this example, fired more intensely for stimuli closer to the category boundary (120/563 at $P < 0.05$, Wilcoxon signed-rank test). A smaller fraction (66/563) showed the opposite tuning, firing at a higher intensity for easy stimuli, those far from the category boundary (Fig. 2e, f).

The observed modulation of firing rate by stimulus difficulty is consistent with previous findings that the response of many OFC neurons correlates with the expected values associated with reward predictive cues[7–10]. Surprisingly, however, when we compared correct and incorrect choices for the same stimulus (for example, the 68/32 mixture), we found that many neurons showed different firing rates even before the outcome was delivered. Figure 3a, b shows an example of a neuron that tended to fire more when the rat had

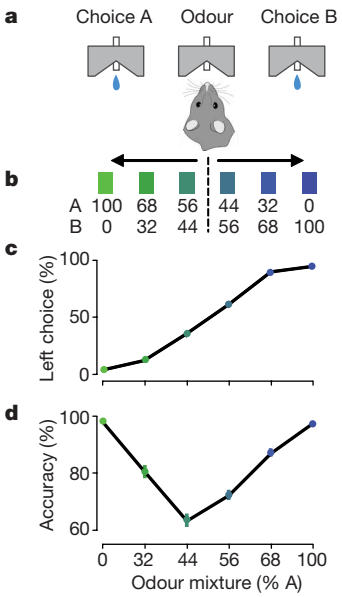


Figure 1 | Odour mixture categorization task. a, Schematic of the behavioural paradigm. To initiate a trial, the rat enters the central odour port and after a pseudorandom delay of 0.2–0.5 s a mixture of odours is delivered. Rats respond by moving to the left or right choice port, where a drop of water is delivered after a 0.3–2 s waiting period for correct choices. **b**, Stimulus design. **c**, Performance of one rat discriminating between mixtures of caproic acid (A) and 1-hexanol (B) in a single session. Error bars (s.e.m.) are hidden by markers. Colours are used to represent odour mixtures, with different blue and green blends representing different odour mixture ratios. **d**, Choice accuracy as a function of odour mixture. Data across three rats are plotted as mean ± s.e.m.

[1]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. [2]Department of Molecular and Cellular Biology and Center for Brain Science, Harvard University, Cambridge, Massachusetts 02138, USA. [3]Allen Institute for Brain Science, Seattle, Washington 98103, USA. [4]Champalimaud Neuroscience Programme, Instituto Gulbenkian de Ciência, 2780-901 Oeiras, Portugal.

committed an error than when it was correct, despite the fact that the outcome was not yet revealed to the subject. The same phenomenon could also be seen as a difference in the average behavioural accuracy when the neuron was firing at high compared to low rates (see Supplementary Fig. 2a). Similar to this example, a large fraction of neurons fired at a higher rate in incorrect trials ('error trials') compared to correct trials within a given stimulus type (46/317 neurons for 56/44 mixtures and 86/563 for 68/32 mixtures at $P < 0.05$, permutation test, Fig. 3d–f; Supplementary Figs 2b and 3c). Interestingly, for easier stimuli the difference in firing rates between correct and error trials was larger (Fig. 3; Supplementary Fig. 3d). A second, smaller population of neurons (21/317 for 56/44 mixtures and 50/563 for 68/32 mixtures at $P < 0.05$, permutation test) had an analogous pattern of activity, but fired more in anticipation of correct rather than incorrect outcomes (Supplementary Fig. 4).

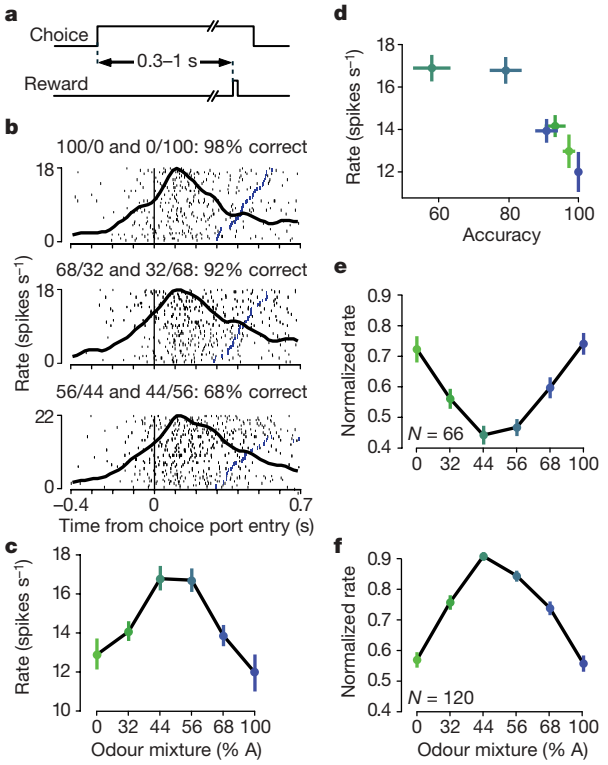


Figure 2 | Graded representation of stimulus difficulty in orbitofrontal cortex. **a**, Timing of outcome anticipation period. Entry into the choice port is recorded using the interruption of the photo-beams within each port. The delivery of water is pseudo-randomly delayed, with the earliest onset varying between 0.3 s and 1 s and the latest offset from 0.8 s to 2 s after entry, according to a uniform distribution with varying parameters in each session. The anticipation period ends at the first possible time of reward delivery, and thus ranges from 0.3 s to 1 s across sessions. Firing rates are calculated either during the initial 0.4 s of the anticipation period or the entire period if it was shorter. **b**, Activity of an example neuronal unit. Raster plots represent neural activity, with each row corresponding to a single trial and each tick mark to a spike. Forty trials are shown in each plot with the post-stimulus time histogram (PSTH) overlaid (smoothed with a Gaussian filter, s.d. = 25 ms). Neural activity is aligned to the timing of entry into the choice port. Blue ticks represent the time of reward delivery. Trials for different stimuli were interleaved in the sessions but grouped into different panels according to stimulus difficulty, with stimuli and performance indicated above. **c**, Mean firing rate of cell in **b** as a function of stimulus identity. Rates are calculated during the outcome anticipation period (0.3 s window beginning at the time of entry into the choice port). Error bars, s.e.m. across trials. **d**, Mean firing rate as function of mean accuracy grouped by stimulus identity. **e**, Mean-normalized firing rate as a function of stimulus identity for the population of neurons with higher firing rates in error trials (Wilcoxon test, $P < 0.05$). **f**, As **e** but for the population of neurons with higher firing rates in correct trials (Wilcoxon test, $P < 0.05$).

These firing patterns appear paradoxical for a prediction made on the basis of overall stimulus–outcome associations. However, reward predictions may be generated by a dynamic learning process based on recent reinforcement history[21–23]. To test this idea, we used a more powerful multiple linear regression model to try to predict the firing rate of a given trial based on the history of recent reward outcomes and other externally observable variables (the stimulus and choice direction). This analysis revealed that although a subset of OFC neurons do carry information about past trial events, these account for a relatively small fraction of the firing rate variance compared to what can be explained by the anticipated current trial outcome (Supplementary Fig. 5; for details see Methods). Therefore, the signals we observed in OFC neurons could not be readily explained as reward expectancy based on either a simple average stimulus–reward association or more complex predictions based on reinforcement history.

In principle, the probability of a correct trial outcome could be estimated based on a subjective measure of confidence about the decision. We hypothesized that a useful confidence metric could be calculated by measuring the reliability and consistency of the values

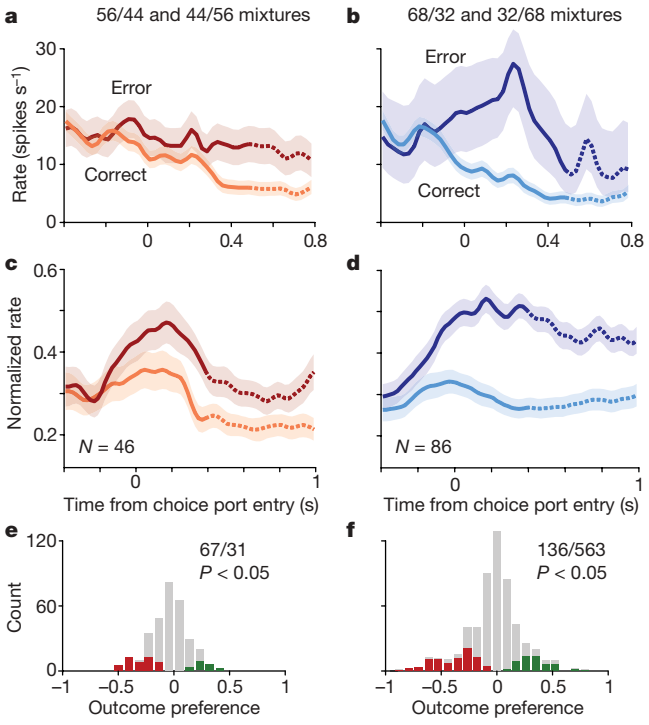


Figure 3 | Orbitofrontal neurons anticipate trial outcome. **a**, **b**, Firing rate of a single neuron aligned to the time of entry into the choice port. Trials are grouped by stimulus difficulty (**a**, 44/56 and 56/44 odour mixture ratio; **b**, 32/68 and 68/32) and trial outcome (correct, orange and cyan; error, red and blue). Shading represents s.e.m.; note there are few 68/32 error trials. Only activity occurring before the onset of water delivery and choice port exit is averaged into the PSTH. After the outcome anticipation period (0.5 s in this session) the PSTH curves are dashed, signifying a time period when in some trials rats experienced reward delivery, although post-reward firing is never actually included. Note that the separation between correct and error trials begins before entry into the choice port but after the animal leaves the odour sampling port. **c**, **d**, Mean-normalized firing of negative outcome selective neurons (those with increased firing rate in error trials during the anticipation period) is plotted the same way as **a**, **b**. Shading represents s.e.m. across neurons. Dashed curves as in **a**, **b**. **e**, **f**, Outcome preference for the population of OFC cells during the outcome anticipation period. Outcome preference is calculated using ROC analysis (see Methods). Colour bars represent significant selectivity (permutation test, $P < 0.05$); red indicates neurons with increased firing rates in incorrect ('error') trials (negative outcome selectivity, 46/317 neurons); green indicates neurons with increased firing rates in correct trials (positive outcome selectivity, 22/317 neurons); grey bars, not significant.

of the internal variables that contributed to the decision. To explore this idea, we constructed a simple model for the categorization task based on the comparison of the perceived stimulus value and the recalled category boundary (Fig. 4a; see Methods for details). In this model, the choice depends on whether the stimulus sample, s_i, is smaller or larger than the category boundary, b_i. This comparison yielded an average choice function similar to that observed behaviourally (Fig. 4b; compare Fig. 1c). To estimate the confidence about this choice, we propose to measure the quality of the evidence in this model using the distance between the stimulus and memory samples, $d_i = |s_i - b_i|$; the larger the distance, the more reliable should be the decision. We found that after a simple transformation, d_i can indeed provide a veridical prediction of the likelihood of a successful outcome, 'decision confidence', $\delta_i = f(d_i)$, or the likelihood of a failure, 'decision uncertainty', $\sigma_i = 1 - \delta_i$ (Fig. 4c). Similar algorithms can also yield useful confidence estimates in other decision models. For example, in a two-alternative 'race' model, an instance of a class of models based on the accumulation of evidence[4–6], decision confidence can be calculated from the difference between two decision variables at the time a decision is reached (Supplementary Fig. 6; Supplementary Information). These modelling results demonstrate that confidence estimates derived solely from the decision variables in the current trial can provide good estimates of the expected decision outcome across trials.

We next looked for specific predictions—patterns of firing rates—that would arise from theoretical confidence estimates. We noticed that, when plotted as a function of stimulus type and trial outcome, decision uncertainty, σ_i, shows a characteristic and somewhat counterintuitive pattern, namely opposing V-shaped curves for correct and error choices (Fig. 4d): (1) for correct choices, σ_i decreases with distance from the category boundary; (2) for a given stimulus, error trials are associated with higher σ_i than correct trials; (3) the difference in σ_i for error and correct trials increases as the stimulus becomes easier. These patterns are robust to model details and do not depend on the relative contributions of stimulus versus memory noise or on the precise choice of the transform function, f (Supplementary Fig. 7). In addition, the same pattern of confidence estimates are produced by decision models based on integration of evidence (Supplementary Fig. 6).

The dependence of OFC neuronal activity on stimulus type and trial outcome closely matched the predictions of confidence estimates derived from decision models (Fig. 4e–h). First, individual OFC neurons showed the predicted dependence on the distance of the stimulus to the category boundary as well as the predicted difference between correct and error trials (Fig. 4e). A similar pattern held at the population level (Fig. 4g, 133/563 negatively-tuned neurons, all stimuli pooled at $P < 0.05$, permutation test; see also Supplementary Figs 3, 8). These patterns were qualitatively different from those expected from left/right modulation of stimulus selectivity (Supplementary Fig. 3). Second, the probability of correct trial outcome varied with the firing rate of individual neurons (Fig. 4f), and at the population level (Fig. 4h), as predicted (Fig. 4c). This analysis also showed that the highest firing rates were associated with near chance performance (50% reward probability), as expected if these neurons signalled lack of confidence rather than incorrect performance (0% reward probability; see Methods for details). The opposite patterns held for the positive outcome selective OFC population (105/563 neurons for all stimuli pooled at $P < 0.05$, permutation test; Supplementary Fig. 4).

It is possible for the experimenter observing OFC neurons to predict individual trial outcomes, but can rats use such information behaviourally? We tested the ability of rats to provide a behavioural report of confidence using a modified version of the task in which we encouraged rats to give up waiting for uncertain rewards by increasing the delay to reward delivery and permitting subjects to reinitiate a

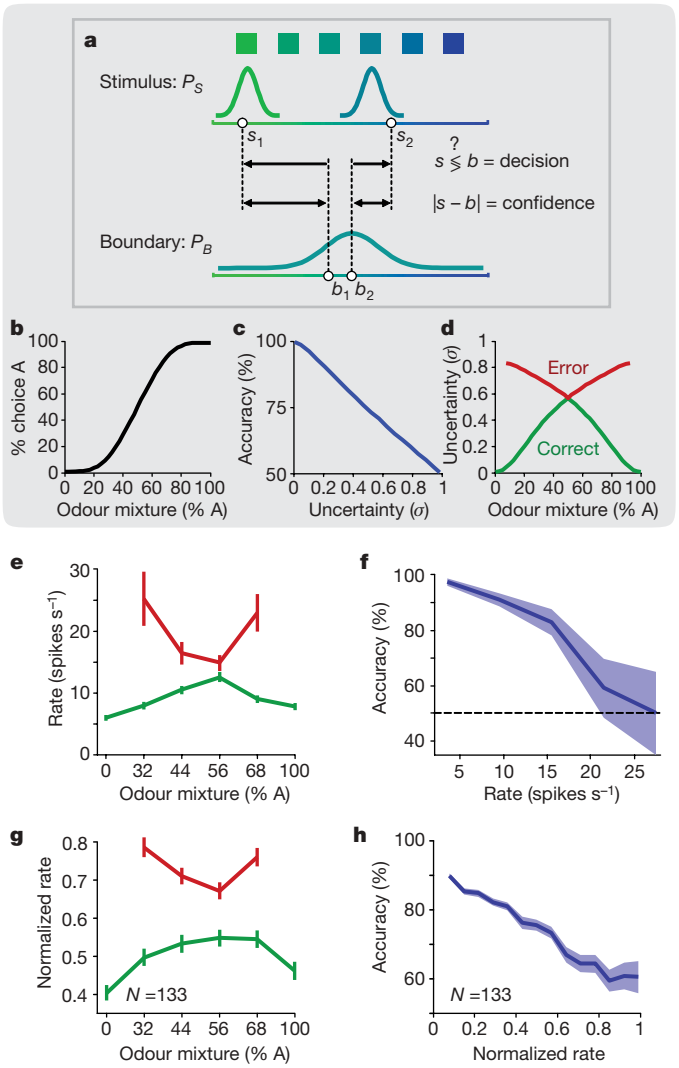


Figure 4 | Confidence estimation in a decision model and by OFC neurons. **a**, Schematic of a model for category decisions. Each odour mixture stimulus, as well as the memory for the category boundary, is encoded as a distribution of values. In each trial a stimulus, s_i, and memory of the boundary, b_i, are drawn from their respective distributions. A choice is calculated by comparing the two samples ($s_i < b_i$), and a confidence value is estimated by calculating their distance ($|s_i - b_i|$). Incorrect choices result from noise, represented in the model by the width of the stimulus and category boundary distributions. See Methods for details. **b**, Example psychometric function of the model, replicating the high choice accuracy of rats for pure odours and decreased accuracy for mixtures near the imposed the category boundary. **c**, Mean accuracy of model choices as a function of decision uncertainty. The uncertainty estimate, σ, is transformed from the distance between the stimulus and boundary samples ($\sigma_i = 1 - \tanh(|s_i - b_i|)$), see Methods). **d**, Mean decision uncertainty estimates generated by the model as a function of stimulus and trial outcome. Note that the model (or a subject) has access only to a stimulus sample and not the stimulus type (for example, 56/44) (see Supplementary Information for an explanation of the pattern of uncertainty estimates.). **e**, Firing rate of an example neuron (same unit as Fig. 3a, b) during the outcome anticipation period as a function of odour stimulus and trial outcome. Error bars are s.e.m. across trials. **f**, Mean choice accuracy as a function of the firing rate for the same unit in **e**. Firing rates were binned and the mean accuracy was calculated for each range of firing rates. Error bars represent standard errors based on the binomial distribution of outcomes. **g**, Mean normalized firing rate of negative outcome selective population (negative outcome preference index across trials with all stimuli pooled at $P < 0.05$, permutation test) during the anticipation period. **h**, Mean accuracy as a function of the firing rate for the same neuron population as in **g**. Firing rates were binned for individual neurons and the mean accuracy was calculated for each range of firing rates. These curves were normalized to a maximal firing rate of 1 and averaged. Error bars represent s.e.m. across neurons.

trial (Fig. 5a). While waiting at the choice port, the decision whether to stay and wait for a possible reward or to go and reinitiate the trial could benefit from an estimate of the confidence in the original decision. Indeed, we found that rats preferentially aborted uncertain trials. Like the neural responses in OFC, these response patterns closely agreed with the predictions of the decision confidence model (Figs 5b, c and 4d). Therefore rats not only show a neural correlate of decision confidence but they can use such information in subsequent decisions to guide adaptive behaviour.

The patterns of neural activity and behaviour we observed suggest that when a decision is made the brain not only makes a choice but also generates an evaluation about the quality of evidence that contributed to the decision. We liken this to the way *P* values are assigned to statistical statements. Our interpretation of the data rests on two results: first, we defined a mechanism for computing confidence in simple decision models and showed that this produced a close fit to a non-trivial pattern of neural and behavioural data; second, we ruled out alternative models for the data, principally ones based on learning. Confidence estimates based on internal decision variables provide useful information that is not readily gained by observing the past relationships between externally observable stimulus, response and outcome variables. Intuitively, this is possible because the observable result of a decision, the choice, is only a partial distillation of the information entering the internal decision process. Computing decision confidence essentially requires calculating how 'close a call' was the choice or how well the evidence was in agreement. When decision 'noise' arises from sources internal to the brain, this process is inherently subjective (accessible only to the subject). More formally, decision confidence can be expressed as the variance measured across the set of decision variables contributing to a single trial (see Supplementary Information). Two different classes of decision model yielded very similar results, suggesting a degree of generality to our description. Nevertheless, it will be important to examine the properties of other methods for estimating confidence.

A variety of results suggests that a key function of OFC is to generate reward predictions based on stimulus–reward associations[7–10]. Our data support and extend this idea by showing that OFC neurons signal outcome predictions derived from a different source, specifically, from internal variables contributing to a perceptual decision on a given trial. In addition to predicting expected rewards, OFC has also been implicated in signalling outcome risk or variance[16–20]. Because in a two-alternative psychophysical decision task the expected reward and its variance are closely related, our data are consistent with both functions and further experiments will be needed to distinguish between these alternatives. It also remains to be determined whether OFC neurons drive the reinitiation behaviour displayed by rats (Fig. 5) or other behaviours contingent on confidence estimates. Indeed, decision confidence signals could be useful for a variety of functions, including controlling exploration[24,25], modulating learning rates[26] and focusing attention[27,28].

Bayesian theory suggests that uncertainty estimates must be incorporated into neural computations for optimal behaviour[29]. Humans and other primates clearly have the ability to assess and act on the degree of uncertainty or confidence in their beliefs about the world[1,11,30], but it has been argued that this might be a sophisticated 'metacognitive' capacity requiring self-awareness[13,14] and a neural architecture specific to primates[11]. Our results show that rodents possess the ability to act on their degree of belief in a decision[12] and demonstrate that estimating the confidence in a choice is little more complex than calculating the choice itself. It is likely that confidence estimates for memories or other beliefs[11,30] could be derived in an analogous fashion. We suggest that the computation of subjective confidence may be a core component of decision-making that, like subjective value signals[7–10,21–23], is important to a wide range of behaviours and their neural substrates.

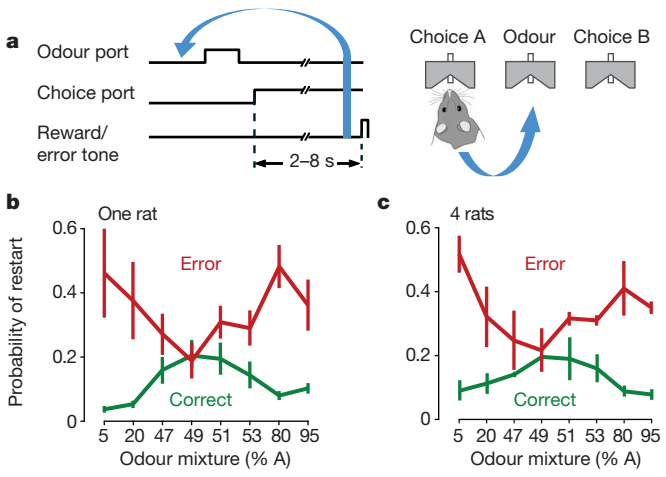


Figure 5 | Behavioural use of decision confidence. **a**, Schematic of the reinitiation task. Reward delivery was pseudo-randomly delayed between 2 and 8 s (uniform distribution) after the rat's choice was registered. Incorrect choices were signalled with an error tone delivered at the end of the 8 s delay. There was a minimum delay of 2 s from the time of the choice before rats could initiate a new trial. **b**, Probability of reinitiation for a single rat plotted as a function of odour stimulus and trial outcome. Error bars represent s.e.m. across trials. Entry into the odour port within 2 s of aborting was considered a reinitiation. **c**, Mean probability of reinitiation for 4 rats as a function of odour stimulus and trial outcome. Error bars represent s.e.m. across rats.

METHODS SUMMARY

Male Long-Evans hooded rats were trained to perform an odour categorization task for water reward. Behavioural testing was controlled by custom software written in Matlab (Mathworks) using data acquisition hardware (National Instruments) to record the port signals and control the valves of the olfactometer and water-delivery[15].

Rats were implanted with custom-made microdrives in the left orbitofrontal cortex (3.5 mm anterior to bregma and 2.5 mm lateral to midline). Extracellular recordings were obtained with six independently movable tetrodes using the Cheetah system (Neuralynx) and single units were isolated by manually clustering spike features with MClust (A. D. Redish).

We focused our analysis on the 'reward anticipation period' while rats remained at one of the choice ports. This excluded spikes that occurred during or after water valve actuation on correct trials; on error trials, no feedback was present. To determine how well neural activity predicted the upcoming outcome (reward/no reward), we used receiver operating characteristics (ROC) analysis to calculate an outcome preference index (OP) that measures how well an ideal observer can predict the outcome from the knowledge of the firing rate from trial to trial. This index varies from −1 to 1 with the sign denoting whether a neuron fires more for rewarded (correct, +) or unrewarded (error, −) decisions:

$$\mathrm{OP} = 2(\mathrm{ROC}_{\mathrm{area}} - 0.5); \quad \mathrm{ROC}_{\mathrm{area}} = \int_0^\infty P(f_{\mathrm{correct}} = f)P(f_{\mathrm{error}} < f)\mathrm{d}f \quad \text{where} \quad f_{\mathrm{correct}}$$

and f_{error} refer to the distribution of firing rates during the reward anticipation period in correct and error trials respectively. Statistical significance was evaluated using a permutation test, where trial order was pseudo-randomly shuffled 200 times to yield a *P* value.

All procedures involving animals were carried out in accordance with National Institutes of Health standards and were approved by the Cold Spring Harbor Laboratory Institutional Animal Care and Use Committee.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 28 February; accepted 26 June 2008.
Published online 10 August 2008.

1. Kahneman, D., Slovic, P. & Tversky, A. *Judgment under Uncertainty: Heuristics and Biases* (Cambridge Univ. Press, 1982).
2. Glimcher, P. W. *Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics* (MIT Press, 2003).
3. Kim, J. N. & Shadlen, M. N. Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nature Neurosci.* **2**, 176–185 (1999).

4. Bogacz, R. *et al.* The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* **113,** 700–765 (2006).
5. Mazurek, M. E., Roitman, J. D., Ditterich, J. & Shadlen, M. N. A role for neural integrators in perceptual decision making. *Cereb. Cortex* **13,** 1257–1269 (2003).
6. Ratcliff, R. & Smith, P. L. A comparison of sequential sampling models for two-choice reaction time. *Psychol. Rev.* **111,** 333–367 (2004).
7. Schoenbaum, G., Chiba, A. A. & Gallagher, M. Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nature Neurosci.* **1,** 155–159 (1998).
8. Tremblay, L. & Schultz, W. Relative reward preference in primate orbitofrontal cortex. *Nature* **398,** 704–708 (1999).
9. Padoa-Schioppa, C. & Assad, J. A. Neurons in the orbitofrontal cortex encode economic value. *Nature* **441,** 223–226 (2006).
10. Wallis, J. D. Orbitofrontal cortex and its contribution to decision-making. *Annu. Rev. Neurosci.* **30,** 31–56 (2007).
11. Smith, J. D., Shields, W. E. & Washburn, D. A. The comparative psychology of uncertainty monitoring and metacognition. *Behav. Brain Sci.* 26, 317–339 340–373 (2003).
12. Foote, A. L. & Crystal, J. D. Metacognition in the rat. *Curr. Biol.* **17,** 551–555 (2007).
13. Persaud, N., McLeod, P. & Cowey, A. Post-decision wagering objectively measures awareness. *Nature Neurosci.* **10,** 257–261 (2007).
14. Koch, C. & Preuschoff, K. Betting the house on consciousness. *Nature Neurosci.* **10,** 140–141 (2007).
15. Uchida, N. & Mainen, Z. F. Speed and accuracy of olfactory discrimination in the rat. *Nature Neurosci.* **6,** 1224–1229 (2003).
16. Bechara, A., Damasio, H., Tranel, D. & Damasio, A. R. Deciding advantageously before knowing the advantageous strategy. *Science* **275,** 1293–1295 (1997).
17. Critchley, H. D., Mathias, C. J. & Dolan, R. J. Neural activity in the human brain relating to uncertainty and arousal during anticipation. *Neuron* **29,** 537–545 (2001).
18. Grinband, J., Hirsch, J. & Ferrera, V. P. A neural representation of categorization uncertainty in the human brain. *Neuron* **49,** 757–763 (2006).
19. Hsu, M. *et al.* Neural systems responding to degrees of uncertainty in human decision-making. *Science* **310,** 1680–1683 (2005).
20. Tobler, P. N., O'Doherty, J. P., Dolan, R. J. & Schultz, W. Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. *J. Neurophysiol.* **97,** 1621–1632 (2007).
21. Barraclough, D. J., Conroy, M. L. & Lee, D. Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neurosci.* **7,** 404–410 (2004).
22. Sugrue, L. P., Corrado, G. S. & Newsome, W. T. Matching behavior and the representation of value in the parietal cortex. *Science* **304,** 1782–1787 (2004).
23. Lau, B. & Glimcher, P. W. Dynamic response-by-response models of matching behavior in rhesus monkeys. *J. Exp. Anal. Behav.* **84,** 555–579 (2005).
24. Stephens, D. W. & Krebs, J. R. *Foraging Theory* (Princeton Univ. Press, 1986).
25. Behrens, T. E., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. Learning the value of information in an uncertain world. *Nature Neurosci.* **10,** 1214–1221 (2007).
26. Yu, A. J. & Dayan, P. Uncertainty, neuromodulation, and attention. *Neuron* **46,** 681–692 (2005).
27. Dayan, P., Kakade, S. & Montague, P. R. Learning and selective attention. *Nature Neurosci.* **3** (Suppl), 1218–1223 (2000).
28. Luck, S. J., Hillyard, S. A., Mouloua, M. & Hawkins, H. L. Mechanisms of visual-spatial attention: Resource allocation or uncertainty reduction? *J. Exp. Psychol. Hum. Percept. Perform.* **22,** 725–737 (1996).
29. Knill, D. C. & Pouget, A. The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends Neurosci.* **27,** 712–719 (2004).
30. Hampton, R. R. Rhesus monkeys know when they remember. *Proc. Natl Acad. Sci. USA* **98,** 5359–5362 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Paton, A. Pouget, S. Raghavachari, G. Turner and members of the Mainen laboratory for comments on the manuscript. Support was provided by the National Institutes of Health (NIDCD) (Z.F.M.), the Center for the Neural Mechanisms of Cognition at Cold Spring Harbor Laboratory (Z.F.M.), and the Swartz Foundation (A.K., N.U., Z.F.M).

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.K. (kepecs@cshl.edu) or Z.F.M. (zmainen@igc.gulbenkian.pt).

METHODS

Here we describe the behavioural and physiological methods used in this study and explain the analyses presented in the main text.

Behavioural task. The behavioural box contains a panel of three ports: the central port for odour delivery ('odour port'), and two ports on each side ('choice ports') for water delivery (Fig. 1a). Entry and exit from the ports was detected based on an infrared photo-beam located inside each port. Odours were mixed with pure air to produce a 1:20 dilution at a flow rate of 1 l min^{-1} using a custom-built olfactometer[15].

Rats self-initiated each experimental trial by introducing their snout into a central port where odour was delivered (Fig. 1a). After a variable delay, drawn from a uniform random distribution of 0.2–0.5 s, a binary mixture of two pure odorants, caproic acid and 1-hexanol, was delivered at one of 4–6 concentration ratios (100/0, 68/32, 56/44, 44/56, 32/68, 0/100; Fig. 1b) in pseudorandom order within a session. After a variable odour sampling time up to 1 s, rats responded by withdrawing from the central port, which terminated the delivery of odour, and moved to the left or right choice port (Fig. 1a). Choices were rewarded according to the dominant component of the mixture, that is, at the left port for mixtures A/B < 50/50 and at the right port for A/B > 50/50 (Fig. 1b). We introduced a variable reward delay period after entry into the choice port. For correct choices, reward was delivered between at least 0.3 s after entry into the choice port and sometimes up to 2 s (in individual sessions the delays were uniformly distributed with the onset ranging from 0.3–0.8 s and the offset to 1–2 s). Outcome selectivity calculations used firing rates calculated over the first 0.4 s of the reward anticipation period. In a few sessions the reward anticipation was 0.3 s (e.g. Fig. 2c, d); in those sessions the entire reward anticipation period was used.

This task allowed us to control the distance of each stimulus to the category boundary and hence systematically manipulate the difficulty of individual categorization problems (Fig. 1d). Intuitively, this task is analogous to categorizing colours along a continuous spectrum (for example, blue/green, Fig. 1b). For colour blends in the middle, the answer depends on a semi-arbitrary convention of colour category boundaries. Similarly, our training protocol enforced the 50/50 odour category boundary, which is semi-arbitrary, as the pure odours do not have equal intensity.

Reinitiation task. In this version of the task, the delay to reward was increased to between 2 and 8 s (uniform random distribution). Errors were signalled with an auditory beep at 8 s and punished with an additional 4 s time-out. After a 2 s mandatory wait from the entry into a choice port and before water or auditory feedback was provided, subjects were allowed to abort trials by exiting the water port. Entry into the odour port within 2 s of aborting was considered as 'reinitiation'. The stimulus ensemble consisted of 75% easy (95/5, 80/20 mixtures: 92 ± 4% accuracy, s.e.m across rats) and 25% difficult (53/47, 51/49 mixtures: 55 ± 2% accuracy) stimuli so that rats could expect to encounter an easier stimulus after reinitiating a new trial. The expectation of a rat to receive reward by staying at the choice port should be proportional to its confidence about the first choice (Fig. 4d) while the expectation to receive reward by reinitiating a new trial should be fixed (because the new stimulus is not predictable). Therefore the relative value of reinitiating is predicted to increase as confidence drops, with approximately the same dependence on stimulus and outcome as given by the model (Fig. 4c). The exact value depends on the actual delays and the subject's temporal discounting function.

Neural data collection and analysis. Rats were implanted with custom-made microdrives in the left orbitofrontal cortex (3.5 mm anterior to bregma and 2.5 mm lateral to midline) as described previously[31] (Supplementary Fig. 1). Extracellular recordings were obtained using six independently adjustable tetrodes for recording. Electrodes were advanced each recording day to sample an independent population of cells across sessions. The placement of electrodes was estimated by depth and confirmed with histology. Neural and behavioural data were synchronized by acquiring time-stamps from the behavioural system along with the electrophysiological signals. Data analysis was performed using Matlab (Mathworks).

For Fig. 2e, f, confidence-modulated neurons were selected by performing a non-parametric, Wilcoxon signed-rank test on firing rates during the reward anticipation period for correct versus error trials. Neurons with significant ($P < 0.05$) firing rate differences were separated into two populations based on whether their mean firing rate was higher for correct or error trials. We then plotted the maximum normalized firing rate averaged for each neural population as a function of stimulus mixture ratio. We used this selection criterion because by not using information about the stimulus it does not impose a specific shape on the tuning curves. Other selection criteria, such as significant rate-accuracy correlations (for example, Fig. 2d), yielded similar results.

Multiple linear regression analysis. We considered the possibility that a prediction of upcoming trial outcome might be made on the basis of recent reward history[32–35] and other observable task variables. For example, if the average performance fluctuated due to changes in attention or motivation and OFC neurons tracked the recent history of trial outcomes, it could lead to a differential prediction of correct versus error trials when averaged over the entire session. In this scenario, outcome selectivity would arise because the present trial's expected outcome is correlated with the recent trials' outcomes. Although we did not observe prominent performance fluctuations, we wanted to test this and related possibilities directly. We used multiple linear regression in an attempt to predict the firing rate of a given trial based on the history of recent reward outcomes and experimental variables (stimulus type and choice direction). Specifically we fitted the firing rates during the reward anticipation period to the following model:

$$\mathrm{RATE}_{t=0} = \alpha_1 S_{t=0} + \alpha_2 C_{t=0} - \sum_{k=0}^{-3} \beta^{\mathrm{L}}_{t=k} O^{\mathrm{L}}_{t=k} - \sum_{k=0}^{-3} \beta^{\mathrm{R}}_{t=k} O^{\mathrm{R}}_{t=k} + \gamma$$

where $S_{t=0}$ represents the stimulus difficulty of the current trial ($t = 0$), which is assumed to be learned through long-term experience with a given stimulus; $C_{t=0}$ represents the choice of sides (left or right, L or R) in the current trial, which is known to influence the firing rate of OFC neurons[31,36]. The variable $O^{\mathrm{SIDE}}_{t=k}$ represents outcomes of the current trial and past three trials ($t = -1, -2, -3$), separated according to the side where the reward was received, again to account for the known selectivity of rodent OFC neurons[31,36]. The coefficients α_1 and α_2 measure the influence of the stimulus difficulty and the choice, $\beta^{\mathrm{L}}_{t=k}$ and $\beta^{\mathrm{R}}_{t=k}$ measure the influence of current and past trial outcomes, and γ captures the mean rate not accounted for by other variables.

The model was fitted using a least-square error criterion with singular value decomposition (SVD). In some cases the problems were ill-conditioned and therefore we also tried ridge regression to obtain more stable solutions. For this analysis, the optimal regularization parameter was chosen by generalized cross-validation[37]. The results of both analyses essentially agreed and therefore we report the results from SVD estimated regression models. The statistical significance of regression coefficients was determined using a permutation test by pseudo-randomly shuffling trial order for the variable of interest[38]. The data were shuffled 1,000 times to yield a P value for the permutation test.

Supplementary Fig. 5a shows the coefficients of this model fit to the neuron shown in Fig. 3a, b. Error bars show standard deviations estimated using leave-one-out-bootstrap[37] and filled circles show significant values at $P < 0.05$ based on a permutation test. This neuron had significant selectivity for the upcoming outcome, $\beta^{\mathrm{L,R}}_{t=0}$, for both choice sides, as well as for the previous outcome, $\beta^{\mathrm{L,R}}_{t=-1}$, to a much smaller degree, while the influence of past outcomes, $\beta^{\mathrm{L,R}}_{t=-2,-3}$, was not significant. Leaving out all past outcomes, $\beta^{\mathrm{L,R}}_{t=-1,-2,-3} = 0$, did not significantly increase the prediction error ($P < 0.05$, permutation test).

This analysis was repeated on the population of 133 neurons (Fig. 4g, h) that were deemed to be negative outcome selective (pooling trials across all stimuli) based on ROC analysis at $P < 0.05$. Supplementary Fig. 5b shows the number of neurons (grey bars) and the mean value of significant regression coefficients (circles, $P > 0.05$). Overall, 121 neurons had significant $\beta^{\mathrm{L,R}}_{t=0}$ coefficients for the current outcome and 70 neurons had significant $\beta^{\mathrm{L,R}}_{t=-1}$ coefficients for the outcome of the previous trial for at least one side. Only four neurons carried past outcome information for at least one side for all three trials back. Comparison of the average value of the significant coefficients for current and past trial outcomes (Supplementary Fig. 5b, circles) shows that even when past trial outcomes had significant coefficients the average value of their weights was only half those for the current trial.

We also performed an analysis to test whether including the history of recent outcomes improves the model fit. To do this, we compared the full model to one in which the coefficients $\beta^{\mathrm{L,R}}_{t=-1,-2,-3}$ were set to zero and used a permutation test to compare the mean prediction errors for the full and reduced model. To obtain a conservative estimate (that is, allow the best chance for inclusion of history terms to increase performance) we did not compensate for the increased complexity of the full model. This analysis showed that for only 12 of 116 neurons did the inclusion of past outcome information, $\beta^{\mathrm{L,R}}_{t=-1,-2,-3}$, significantly reduce the prediction error ($P < 0.05$, permutation test). Moreover, the reduction in error was small, with an average <3% improvement for the full compared to the reduced (current-trial-only) model.

In summary, we conclude that although a subset of OFC neurons do carry information about past outcomes, past trial events account for a relatively small fraction of the firing rate variance compared to what can be explained by the anticipated current trial outcome.

Outcome selectivity analysis. Orbitofrontal cortex is known to signal outcome expectations[39–42], and an apparent prediction of outcome might arise from a combination of stimulus and side selectivity. If firing rates encoded the stimulus difficulty (Fig. 2) and in addition were modulated by the choice side[31,35] one

would expect (1) outcome preference would be inverted across choice sides, and (2) outcome selectivity would be equal or weaker for easier compared to more difficult stimuli. A cartoon of this scenario is shown in Supplementary Fig. 3b, with both an additive and a multiplicative component to the choice side modulation. In contrast, the uncertainty model makes the opposite predictions (Supplementary Fig. 3a and Fig. 4d). Although the average tuning curve for negative outcome selective neurons are similar to what is expected for a representation of uncertainty (Fig. 4g), we wanted to test these predictions on a neuron-by-neuron basis. We used the outcome preference index (OP) to measure whether the firing rates are higher or lower for error trials, and the unsigned version of this measure, the outcome selectivity index (OS = |OP|), to measure whether how strongly firing rates signal different outcomes. These measures are based on signal detection theory and quantify the difference between the firing rates for error and correct trials (see Methods Summary for details). Statistical significance was estimated using a 200-fold permutation test[43] at $P < 0.05$. Note that for these analyses trials had to be subdivided according to several stimulus types and for many neurons there were few error trials available to reliably compare conditions. An insufficient number of error trials can result in either spurious selectivity values due to noise and/or low significance values.

First we tested whether the direction of outcome preference was concordant across sides (that is, regular arrows in Supplementary Fig. 3a, b). We used 310 out of 563 neurons for which there were more than 5 error trials for each of 32/68 and 68/32 stimuli. From these neurons 116 showed outcome selectivity across all stimuli, but only 19 were significantly selective for both 32/68 and 68/32 mixtures when considered separately. 85% (16/19) of neurons had concordant outcome preference values, and the preference values were significantly correlated across sides ($r^2 = 0.66$, $P < 0.05$; Supplementary Fig. 3c). Next we tested whether outcome selectivity was stronger for easy stimuli (32/68 and 68/32 mixtures) compared to more difficult ones (44/56 and 56/44 mixtures; see dashed arrows in Supplementary Fig. 3a, b). Out of 317 neurons with 56/44 trials, 131 were selective across all stimuli but only 23 were significant for both easy and difficult mixtures when considered separately. For 91% (21/23) of these neurons, outcome selectivity was stronger for easier stimuli (Supplementary Fig. 3d). These analyses support the uncertainty model (Supplementary Fig. 3a) and are not consistent with the hypothesis that choice side-modulation of stimulus encoding neurons produces an apparent outcome selectivity (Supplementary Fig. 3b).

Next we conducted an additional analysis to show how well individual neurons conform to the firing patterns expected for decision confidence across the entire recorded OFC population. We used OP to measure whether the firing rates are higher or lower for error versus correct trials across 32/68, 44/56, 56/44 and 68/32 stimuli. In addition, we calculated a stimulus difficulty selectivity index (DI) to measure whether firing rates are higher or lower for correct choices in difficult trials (32/68, 44/56, 56/44 and 68/32 stimuli) compared to easy trials (0/100 and 100/0 stimuli). Again, both measures are derived from the area under the ROC from signal detection theory and statistical significance was estimated using a 200-fold permutation test at $P < 0.05$. Supplementary Fig. 8 shows DI as a function of OP across the entire population. Out of 563 neurons, 83 were significant for both measures, 85 for OP alone, 105 for DI alone and 290 were not significant at $P < 0.05$. The selectivity measures were correlated (CC = 0.75 at $P < 0.05$) across the entire population. This analysis shows that across the population without any preselection there is a good correlation between outcome preference (selectivity for correct/error choices) and stimulus difficulty preference (selectivity for more/less difficult stimuli) as expected for a decision confidence signal.

Interpretation of negative outcome selectivity: error signal or uncertainty? The observed selectivity of neural activity for the upcoming outcome might arise if, after executing a choice, extra sensory or memory information enters decision-making circuits and causes the realization that an error occurred even before obtaining feedback. According to this interpretation the negative outcome selective population of OFC neurons would signal error[44] instead of uncertainty. In contrast, the highest observed firing rates were associated with near chance level performance and not errors (Fig. 4g, f). To test this more rigorously, we asked whether an ideal observer could obtain better performance than the experimental subject if it could switch choices based on the firing rate after the choice and before feedback is provided. In all but one negative outcome selective neuron (1/133), the highest firing rates (top 5% of trials) were associated with chance level performance (within the 95% confidence interval). Therefore negative outcome selectivity does not imply that OFC neurons are actually able to predict error trials but rather that high firing rates predict near chance level performance consistent with an uncertainty signal.

Confidence model. We model the stimulus as the log ratio of the odour mixture with additive Gaussian noise: $s_i = \log\frac{[A]}{[B]} + \eta_{stim}$ in each trial i, where $\eta_{stim} \in N(0,\sigma_{stim})$. The boundary is fixed at 0 with additive noise, $b^i = \eta^{bound}$, where $\eta_{bound} \in N(0,\sigma_{bound})$. The choice is computed by comparing stimulus and boundary, $choice_i = \{left | s_i < b_i; right | s_i \geq b_i\}$ The distance between the stimulus and boundary, $d_i = |s_i - b_i|$, provides an estimate of decision confidence. Other distance metrics, such as Euclidian distance, are also suitable. This distance measure can be calibrated and linearized to produce a veridical estimate of outcome probabilities[45]. We did not attempt to systematically calibrate confidence but found that sigmoid functions provide a good approximation (see also Supplementary Information). Therefore we define 'decision confidence', $\delta_i = f(d_i) = \tanh(d_i)$ and its opposite 'decision uncertainty' as $\sigma_i = 1 - \delta_i$. For the simulations in Fig. 4 we chose the stimulus and boundary noise to be equal, $\sigma_{bound} = \sigma_{stim} = 0.5$, but we note that the results are dependent only on the total noise (sum of the variances) not their relative contribution (see Supplementary Fig. 7). Therefore, the model has a single effective parameter, $\sigma_{noise} = \sqrt{\sigma_{bound}^2 + \sigma_{stim}^2}$, that determines the slope of the psychometric function, leaving no free parameters with respect to confidence estimates (Supplementary Fig. 7).

31. Feierstein, C. E. *et al.* Representation of spatial goals in rat orbitofrontal cortex. *Neuron* **51,** 495–507 (2006).
32. Barraclough, D. J., Conroy, M. L. & Lee, D. Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neurosci.* **7,** 404–410 (2004).
33. Sugrue, L. P., Corrado, G. S. & Newsome, W. T. Matching behavior and the representation of value in the parietal cortex. *Science* **304,** 1782–1787 (2004).
34. Lau, B. & Glimcher, P. W. Dynamic response-by-response models of matching behavior in rhesus monkeys. *J. Exp. Anal. Behav.* **84,** 555–579 (2005).
35. Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. D Cortical substrates for exploratory decisions in humans. *Nature* **441,** 876–879 (2006).
36. Roesch, M. R., Taylor, A. R. & Schoenbaum, G. Encoding of time-discounted rewards in orbitofrontal cortex is independent of value representation. *Neuron* **51,** 509–520 (2006).
37. Hansen, P. C. *Rank-deficient and Discrete Ill-posed Problems: Numerical Aspects of Linear Inversion* (SIAM, 1998).
38. Davison, A. C. & Hinkley, D. V. *Bootstrap Methods and Their Application* (Cambridge Univ. Press, 1997).
39. Hikosaka, K. & Watanabe, M. Delay activity of orbital and lateral prefrontal neurons of the monkey varying with different rewards. *Cereb. Cortex* **10,** 263–271 (2000).
40. Wallis, J. D. & Miller, E. K. Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *Eur. J. Neurosci.* **18,** 2069–2081 (2003).
41. Gottfried, J. A. O'Doherty, J. & Dolan, R.J. Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science* **301,** 1104–1107 (2003).
42. Simmons, J. M. Ravel, S. Shidara, M. & Richmond, B.J. A comparison of reward-contingent neuronal activity in monkey orbitofrontal cortex and ventral striatum: guiding actions toward rewards *Ann. N.Y. Acad. Sci.* **1121,** 376–394 (2007).
43. Efron, B. & Tibshirani, R. *An Introduction to the Bootstrap* (Chapman & Hall, 1993).
44. Laubach, M., Wessberg, J. & Nicolelis, M. A. Cortical ensemble activity increasingly predicts behaviour outcomes during learning of a motor task. *Nature* **405,** 567–571 (2000).
45. Keren, G. On the calibration of probability judgments. Some critical comments and alternative perspectives. *J. Behav. Decis. Making* **10,** 269–278 (1997).

LETTERS

Large recurrent microdeletions associated with schizophrenia

Hreinn Stefansson[1]*, Dan Rujescu[2]*, Sven Cichon[3,4]*, Olli P. H. Pietiläinen[5], Andres Ingason[1], Stacy Steinberg[1], Ragnheidur Fossdal[1], Engilbert Sigurdsson[6], Thordur Sigmundsson[6], Jacobine E. Buizer-Voskamp[7], Thomas Hansen[8,9], Klaus D. Jakobsen[8,9], Pierandrea Muglia[10], Clyde Francks[10], Paul M. Matthews[11], Arnaldur Gylfason[1], Bjarni V. Halldorsson[1], Daniel Gudbjartsson[1], Thorgeir E. Thorgeirsson[1], Asgeir Sigurdsson[1], Adalbjorg Jonasdottir[1], Aslaug Jonasdottir[1], Asgeir Bjornsson[1], Sigurborg Mattiasdottir[1], Thorarinn Blondal[1], Magnus Haraldsson[6], Brynja B. Magnusdottir[6], Ina Giegling[2], Hans-Jürgen Möller[2], Annette Hartmann[2], Kevin V. Shianna[12], Dongliang Ge[12], Anna C. Need[12], Caroline Crombie[13], Gillian Fraser[13], Nicholas Walker[14], Jouko Lonnqvist[15], Jaana Suvisaari[15], Annamarie Tuulio-Henriksson[15], Tiina Paunio[5,15], Timi Toulopoulou[16], Elvira Bramon[16], Marta Di Forti[16], Robin Murray[16], Mirella Ruggeri[17], Evangelos Vassos[16], Sarah Tosato[17], Muriel Walshe[16], Tao Li[16,18], Catalina Vasilescu[3], Thomas W. Mühleisen[3], August G. Wang[19], Henrik Ullum[20], Srdjan Djurovic[21,22], Ingrid Melle[22], Jes Olesen[23], Lambertus A. Kiemeney[24], Barbara Franke[25], GROUP†, Chiara Sabatti[26], Nelson B. Freimer[27], Jeffrey R. Gulcher[1], Unnur Thorsteinsdottir[1], Augustine Kong[1], Ole A. Andreassen[21,22], Roel A. Ophoff[7,27], Alexander Georgi[28], Marcella Rietschel[28], Thomas Werge[8], Hannes Petursson[6], David B. Goldstein[12], Markus M. Nöthen[3,4], Leena Peltonen[5,29,30], David A. Collier[16,18], David St Clair[13] & Kari Stefansson[1,31]

Reduced fecundity, associated with severe mental disorders[1], places negative selection pressure on risk alleles and may explain, in part, why common variants have not been found that confer risk of disorders such as autism[2], schizophrenia[3] and mental retardation[4]. Thus, rare variants may account for a larger fraction of the overall genetic risk than previously assumed. In contrast to rare single nucleotide mutations, rare copy number variations (CNVs) can be detected using genome-wide single nucleotide polymorphism arrays. This has led to the identification of CNVs associated with mental retardation[4,5] and autism[2]. In a genome-wide search for CNVs associating with schizophrenia, we used a population-based sample to identify *de novo* CNVs by analysing 9,878 transmissions from parents to offspring. The 66 *de novo* CNVs identified were tested for association in a sample of 1,433 schizophrenia cases and 33,250 controls. Three deletions at 1q21.1, 15q11.2 and 15q13.3 showing nominal association with schizophrenia in the first sample (phase I) were followed up in a second sample of 3,285 cases and 7,951 controls (phase II). All three deletions significantly associate with schizophrenia and related psychoses in the combined sample. The identification of these rare, recurrent risk variants, having occurred independently in multiple founders and being subject to negative selection, is important in itself. CNV analysis may also point the way to the identification of additional and more prevalent risk variants in genes and pathways involved in schizophrenia.

The approach we used here was to use a large population-based discovery sample to identify *de novo* CNVs, followed by testing for association in a sample of patients with schizophrenia and psychoses (phase I) and finally replicating the most promising variants from phase I in a second larger sample (phase II). The discovery phase, where

[1]CNS Division, deCODE genetics, Sturlugata 8, IS-101 Reykjavík, Iceland. [2]Division of Molecular and Clinical Neurobiology, Department of Psychiatry, Genetics Research Centre, Ludwig-Maximilians-University, Nußbaumstrasse 7, 80336 Munich, Germany. [3]Department of Genomics, Life & Brain Center, University of Bonn, Sigmund-Freud-Strasse 25, D-53127 Bonn, Germany. [4]Institute of Human Genetics, University of Bonn, Wilhelmstrasse 31, D-53111 Bonn, Germany. [5]Department for Molecular Medicine, National Public Health Institute, Biomedicum, Haartmaninkatu 8, 00290 Helsinki, Finland. [6]Department of Psychiatry, National University Hospital, Hringbraut, 101 Reykjavík, Iceland. [7]The Netherlands Department of Medical Genetics and Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG, Utrecht, The Netherlands. [8]Research Institute of Biological Psychiatry, Mental Health Centre Sct. Hans Copenhagen University Hospital, DK-4000 Roskilde, Denmark. [9]Centre for Pharmacogenomics, University of Copenhagen, DK-2200 Copenhagen N, Denmark. [10]Medical Genetics, GlaxoSmithKline R&D, Via A. Fleming 4, 37135 Verona, Italy. [11]Clinical Imaging Centre, Clinical Pharmacology and Discovery Medicine, GlaxoSmithKline, Hammersmith Hospital, London W12 ONN, UK. [12]Institute for Genome Sciences & Policy, Center for Population Genomics & Pharmacogenetics, 4011 GSRB II 103 Research Drive, Duke University, DUMC Box 3471, Durham, North Carolina 27708, USA. [13]Department of Mental Health, University of Aberdeen, Royal Cornhill Hospital, Aberdeen AB25 2ZD, UK. [14]Ravenscraig hospital, Inverkip Road, Greenock PA16 9HA, UK. [15]Department of Mental Health and Addiction, National Public Health Institute, Mannerheimintie 166, FIN-00300 Helsinki, Finland. [16]Division of Psychological Medicine and Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College, London SE5 8AF, UK. [17]Section of Psychiatry and Clinical Psychology, University of Verona, Verona, 37134 Verona, Italy. [18]Psychiatric Laboratory, Department of Psychiatry, West China Hospital, Sichuan University, Chengdu 610041, Sichuan, China. [19]Department of Clinical Immunology, Copenhagen University Hospital, DK-2200 Copenhagen N, Denmark. [20]Mental Health Centre Amager, Copenhagen University Hospital, DK-2300 Copenhagen S, Denmark. [21]Institute of Psychiatry, University of Oslo, PO Box 1130, Blindern, N-0318 Oslo, Norway. [22]Departments of Medical Genetics and Psychiatry, Ulleval University Hospital, Kirkeveien 166, N-0407 Oslo, Norway. [23]Department of Neurology, 57 Nordre Ringvej, Glostrup Hospital, Glostrup, DK-2600 Copenhagen, Denmark. [24]Department of Epidemiology & Biostatistics (133 EPIB)/Department of Urology (659 URO), Radboud University Nijmegen Medical Centre, PO Box 9101, 6500 HB Nijmegen, The Netherlands. [25]Department of Human Genetics, Radboud University Nijmegen Medical Centre, PO Box 9101, 6500 HB Nijmegen, The Netherlands. [26]Departments of Human Genetics and Statistics, UCLA, 695 Charles Young Drive South, Los Angeles, California 90095, USA. [27]UCLA Center for Neurobehavioral Genetics, Charles E. Young Drive South, Los Angeles, California 90024, USA. [28]Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health, University of Heidelberg, J5, D-68159 Mannheim, Germany. [29]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [30]The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [31]University of Iceland, School of Medicine, Sturlugata 8, 101 Reykjavik, Iceland.
*These authors contributed equally to this work.
†Lists of authors and affiliations appear at the end of the paper.

we searched for *de novo* CNVs, enriches for those regions that mutate most often. If the CNVs identified are in very low frequency in the population despite relatively high mutation rate (>1/10,000 meiosis), they are likely to be under negative selection pressure. Such variants may confer risk of disorders that reduce the fecundity of those affected.

To uncover *de novo* CNVs genome-wide we analysed data from a population-based sample (2,160 trios (two parents and one offspring) and 5,558 parent–offspring pairs, none of which was known to have schizophrenia; Supplementary Table 1), providing information on 9,878 transmissions. Of the 66 *de novo* CNVs identified, 23 were flanked by low copy repeats (LCRs) and nine had a LCR flanking only one of the deletion breakpoints. Of the remaining 34 CNVs (not flanked by LCRs), 27 were only found in a single control sample (the discovery trio) out of the 33,250 tested, whereas 18 out of the 23 CNVs flanked by LCRs were found at a higher frequency in the large control sample (Supplementary Table 2).

The 66 CNVs were tested for association in our phase I sample of 1,433 patients with schizophrenia and related psychoses and 33,250 controls from the SGENE consortium (http://www.sgene.eu/). For eight of the 66 CNVs tested, at least one schizophrenia patient carried the CNV (Supplementary Table 3), and for three large deletions, nominal association with schizophrenia and related psychoses was detected (uncorrected *P*-value <0.05, Table 1). The three deletions nominally associating with schizophrenia in the first sample (Table 1) were followed up in up to six samples comprising a total of 3,285 cases and 7,951 controls (Table 2). All three deletions, at 1q21.1, 15q11.2 and 15q13.3, significantly associate with schizophrenia and psychosis in the combined sample with high odds ratio (OR) ($P = 2.9 \times 10^{-5}$, OR = 14.83; $P = 6.0 \times 10^{-4}$, OR = 2.73; and $P = 5.3 \times 10^{-4}$, OR = 11.54, respectively). Removing cases with psychosis, other than 'diagnostic and statistical manual of mental disorders' and 'research diagnostic criteria' defined schizophrenia (in total 147 cases: 39 with unspecified functional psychosis, 86 with schizoaffective disorder, 10 with schizophreniform and 12 with persistent delusional disorders; Supplementary Information), gave comparable results for the 1q21.1 deletion ($P = 2.31 \times 10^{-5}$, OR = 15.44), whereas the association for 15q11.2 and 15q13.3 deletions was no longer significant ($P = 9.57 \times 10^{-4}$, OR = 2.66 and $P = 1.02 \times 10^{-3}$, OR = 11.29, respectively (uncorrected for 66 tests)). Historically, classification schemes tend to group diseases by their signs and symptoms. There is, however, no reason why the phenotypes associating with a particular CNV should be confined to the current nosological boundaries of any single psychiatric disorder. Our findings, in this respect, resemble those from the 16p11.2 deletion[2] and the translocation disrupting the *DISC1* gene in a large Scottish pedigree[6], and support the idea that the same mutation can increase the risk of a broad range of clinical psychopathology. It is therefore worth noting that among the eight controls carrying the 15q13.3 deletion there is one autistic individual (there are samples from 299 autistic individuals among the 39,800 control samples genotyped for this CNV).

Eleven out of the 4,718 cases tested (0.23%) carry the 1q21.1 deletion compared to eight of the 41,199 controls tested (0.02%). In seven of the eleven patients, the deletion spans about 1.35 megabases (Mb) (chromosome 1: 144,943,150–146,293,282). Four cases have a larger form of the deletion (Supplementary Table 4). The larger form contains the shorter form and extends to 144,106,312 Mb, about 2.19 Mb (Fig. 1a and Supplementary Fig. 1). Seven of the eight Icelandic controls have the shorter form of the deletion and one control has the longer form. Previously reported 1q21.1 deletions in two cases of mental retardation[5,7], two autistic individuals[2] and one schizophrenia case[8] are consistent with the shorter form of the deletion.

The 1.35 Mb deleted segment common to both the large and the small form of the 1q21.1 deletion is gene rich (Fig. 1a). The *GJA8* gene has previously been reported as associated with schizophrenia[9]. This gene is located in a repeat region within the boundary of the 1.35 Mb deletion segment and contains no single nucleotide polymorphism (SNP) markers on the HumanHap300 chip. In at least four reports[10–13] the 1q21 locus has been linked to schizophrenia; however, the deletion is rare and therefore unlikely to account for much of the linkage previously reported. Analysis of cells from a case with the 1q21.1 deletion and a case with the reciprocal duplication, using fluorescence *in situ* hybridization analysis (Supplementary Fig. 2), show that other rearrangements, such as chromosomal translocations, are unlikely to be associated with the deletion.

The deletion at 15q11.2 was significant in the combined schizophrenia and related psychosis sample (Table 2). In the combined sample 26 of 4,718 cases (0.55%) carry the deletion compared with 79 of 41,194 controls (0.19%). The deletion spans approximately 470 kb (chromosome 15: 20,306,549–20,777,695) and several genes are deleted (Fig. 1b and Supplementary Fig. 3). A single case with mental retardation and severe speech impairment has previously been reported with the 15q11.2 deletion[14]. Although the region is not imprinted, it is deleted in a minority of cases of Angelman syndrome and Prader–Willi syndrome. Recent analysis shows that Angelman syndrome cases with class I deletions (includes the 15q11.2 deletion) are significantly more likely to meet criteria for autism[15]. Prader–Willi syndrome type I deletions are associated with increased risk of preservative/obsessive compulsive behaviour, deficits in adaptive skills and lower intellectual ability. Thus, the autistic features in Angelman syndrome and the preservative behaviour of Prader–Willi syndrome may arise from deletion of the genes in the proximal portion of the region, the site at the breakpoints of the chromosome 15 deletions found in the current study. The gene in the 15q11.2 deletion region that is most likely to be responsible for both the autistic and obsessive compulsive features observed in Angelman syndrome and Prader–Willi syndrome with class one deletions, and the schizophrenia phenotype in this study, is *CYFIP1* (Fig. 1c). CYFIP1 interacts with fragile X mental retardation protein (FMRP) as well as with the Rho GTPase Rac1, which is involved in regulating axonal and dendritic outgrowth and the development and maintenance of neuronal structures. Over 30% of children with fragile X syndrome meet criteria for autism[16], with highest rates observed in cases with Prader–Willi features without the deletion on 15q. Notably, the fragile X mutation results in a reduction in expression levels of the *CYFIP1* gene[17], and fragile X syndrome

Table 1 | Nominal association of deletions at 1q21.1, 15q11.2 and 15q13.3 with schizophrenia and related psychoses in the phase I sample

Locus	Chromosome 1: 144.94–146.29 (Mb)		Chromosome 15: 20.31–20.78 (Mb)		Chromosome 15: 28.72–30.30 (Mb)	
	Cases	Controls	Cases	Controls	Cases	Controls
Iceland	1 of 646	8 of 32,442	4 of 646	58 of 32,442	1 of 646	7 of 32,442
Scotland	2 of 211	0 of 229	2 of 211	0 of 229	1 of 211	0 of 229
Germany	1 of 195	0 of 192	3 of 195	0 of 192	1 of 195	0 of 192
England	0 of 105	0 of 96	1 of 105	0 of 96	0 of 105	0 of 96
Italy	0 of 85	0 of 91	0 of 85	0 of 91	0 of 85	0 of 91
Finland	0 of 191	0 of 200	0 of 191	1 of 200	0 of 191	0 of 200
OR		8.68 (1.02, 49.76)		3.90 (1.42, 9.37)		8.94 (0.79, 58.15)
P-value		0.024		0.007		0.040

Three deletions show nominal association with schizophrenia and related psychoses in the first sample of 1,433 patients and 33,250 controls. These deletions are large: the 1q21 deletion spans approximately 1.38 Mb, the one on 15q11.2 approximately 0.47 Mb and the one on 15q13.3 approximately 1.57 Mb. *P*-values (uncorrected for the 66 tests) are from the exact Cochran–Mantel–Haenszel test and are two-sided. Coordinates are based on Build 36 assembly of the human genome. 95% confidence intervals are given within brackets.

Table 2 | Significant association of deletions at 1q21.1, 15q11.2 and 15q13.3 with schizophrenia and related psychoses in the combined samples

Locus	Chromosome 1: 144.94–146.29 (Mb)		Chromosome 15: 20.31–20.78 (Mb)		Chromosome 15: 28.72–30.30 (Mb)	
	Cases	Controls	Cases	Controls	Cases	Controls
Germany	2 of 911	0 of 1,297	3 of 911	4 of 1,297	0 of 911	0 of 1,297
Scotland	2 of 451	0 of 441	5 of 451	1 of 441	0 of 451	0 of 441
The Netherlands	0 of 806	0 of 4,039	4 of 806	12 of 4,039	3 of 806	1 of 4,039
Norway	0 of 237	0 of 272	0 of 237	0 of 272	1 of 237	0 of 272
Denmark*	3 of 442	0 of 1,437	4 of 442	3 of 1,432	0 of 375	0 of 501
China*	0 of 438	0 of 463	0 of 438	0 of 463	NA	NA
Phase II						
OR		∞ (2.85, ∞)		2.18 (1.01, 4.60)		16.47 (1.52, 833.38)
P-value		5.6×10^{-4}		0.032		7.9×10^{-3}
Phase I and II						
OR		14.83 (3.55, 60.40)		2.73 (1.50, 4.89)		11.54 (2.53, 49.58)
P-value		2.9×10^{-5}		6.0×10^{-4}		5.3×10^{-4}

The three deletions nominally significant in phase I were tested for association in follow up samples from Germany, Scotland, The Netherlands, Denmark, Norway and China. All three deletions associate with schizophrenia and related psychoses in the combined phase I and II samples (the multiple testing significance threshold is $0.05/66 = 7.6 \times 10^{-4}$). *P*-values in the table (uncorrected for the 66 tests) are from the exact Cochran–Mantel–Haenszel test and are two-sided. Coordinates are based on Build 36 assembly of the human genome. 95% confidence intervals are given within brackets. NA, not analysed.

*Samples were measured using Taqman assays. Samples with CNVs identified by measuring gene dosage by a Taqman assay were verified and confirmed by genotyping the respective samples using the HumanCNV370 chip. A limited amount of DNA was available for genotyping the Chinese samples.

behavioural abnormalities resemble features of schizophrenia. Fragile X syndrome is caused by the complete loss of function of FMRP, whereas the hemizygous deletion of *CYFIP1* would only cause partial disturbance of FMRP function, in which case an effect similar to that observed in fragile X in females and obligate carriers might be expected. These women have attentional deficit and extreme shyness and anxiety, and they may also present with psychiatric disturbances of which psychotic behaviour is the most frequent[18,19].

The 15q13.3 deletion is also significantly associated with schizophrenia and related psychoses in the combined samples (Table 2). A total of 7 of 4,213 cases (0.17%) carry the deletion and 8 of 39,800 controls (0.02%). One of several affected genes (Fig. 1c and

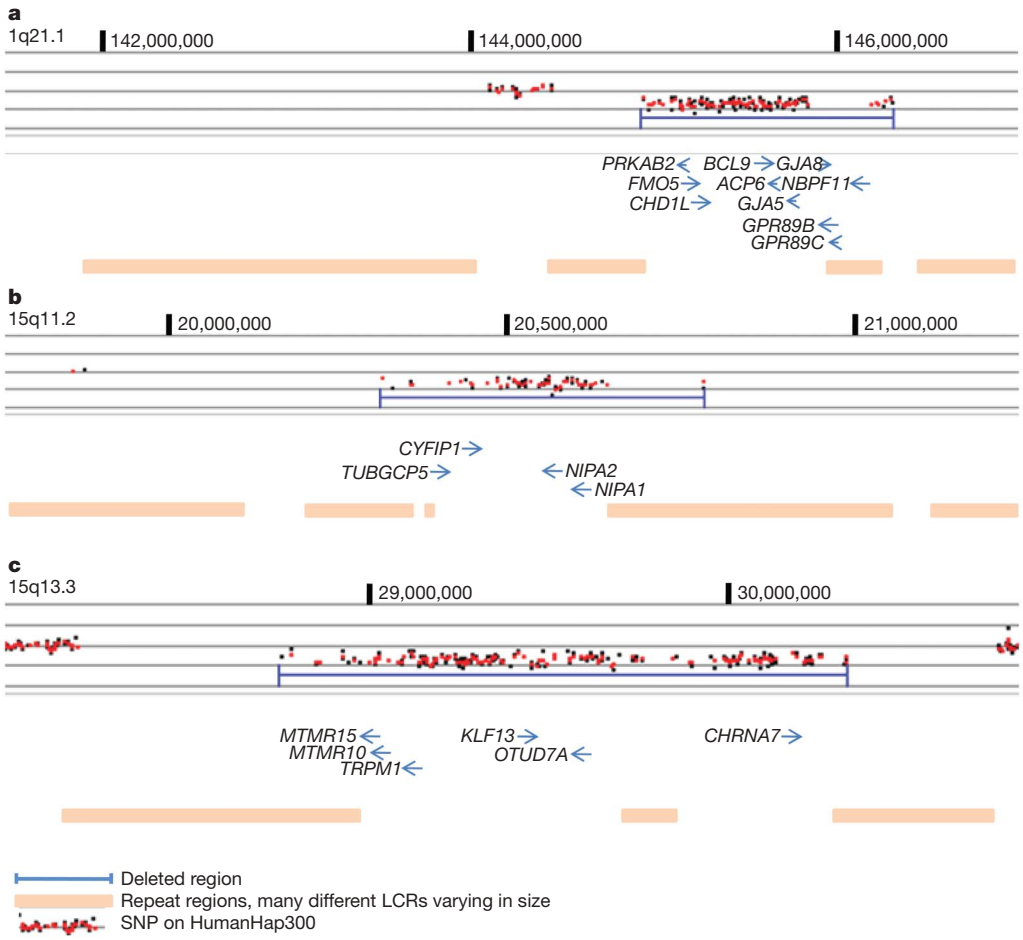


Figure 1 | The genomic architecture of the 1q21.1, 15q11.2 and 15q13.3 deletions. **a**, DosageMiner output showing the shorter form of the 1q21 deletion (marked in blue). Ninety-nine SNPs on the HumanHap300 chip are affected by the deletion which spans 1.38 Mb. **b**, DosageMiner output showing the 15q11.2 deletion (marked in blue). Fifty-four SNPs on the HumanHap300 chip are affected by the deletion which spans 470 kb. **c**, DosageMiner output showing the 15q13.3 deletion (marked in blue). One-hundred-and-sixty-six SNPs on the HumanHap300 chip are affected by the deletion which spans 1.57 Mb. Genes affected by the deletions are shown (coordinates are based on Build 36 of the human genome and positions of genes derived from the UCSC genome browser). LCRs flank all three deletions (Supplementary Figs 1, 3 and 4).

Supplementary Fig. 4), the α7 nicotinic receptor gene (*CHRNA7*), is targeted to axons by neuregulin 1[20], and has been implicated in schizophrenia[21] and also in mental retardation[22]. Mice lacking the α7 subunit of the neural nicotinic receptor show a minor impairment in the matching-to-place task of the Morris water maze, taking longer to find the hidden platform than their wild-type controls. This suggests a role for CHRNA7 in working/episodic memory and a potential role for CHRNA7 in schizophrenia and its endophenotypes[23].

On the HumanHap300 array, 99 SNPs are affected by the deletion on 1q21.1, 54 by the 15q11.2 deletion and 166 by the 15q13.3 deletion (Supplementary Tables 7–9). Significant association was not found with schizophrenia and SNPs at the three deletion loci. However, rare variants at these loci might still associate with schizophrenia as they are not tagged by markers on the HumanHap300 chip. Finding such variants probably requires re-sequencing of the deleted interval in a large sample of cases and testing identified variants for enrichment in schizophrenia.

From available records, we see that cases carrying the 1q21.1, 15q11.2 and 15q13.3 deletions have clinical response rates to neuroleptics that are comparable to the general schizophrenic patient population. Family history of schizophrenia in close relatives is also comparable to other patients with schizophrenia in our sample (although these affected relatives are not available for genotyping) and there is no obvious sex bias, as both males and females carrying the deletions are affected. Assessment of cognitive abilities was only available for a fraction of the cases with deletions. None of the cases carrying the three deletions was known to be mentally retarded; however, three cases carrying the 1q21.1 deletion had learning disabilities and two controls had dyslexia (Supplementary Tables 4–6).

The frequency of the deletions identified here is comparable to the frequency of the velo-cardio-facial syndrome (VCFS) deletion on 22q11, previously shown to associate with schizophrenia[24,25]. The large VCFS deletion was present in 8 out of 3,838 cases tested (0.2%) (Icelandic ($n = 1$), Scottish ($n = 5$), Dutch ($n = 1$) and German (Bonn, $n = 1$)) but was absent in 39,299 controls ($P = 4.2 \times 10^{-5}$, OR $= \infty$).

The CNVs associating with schizophrenia on chromosome 1q21.1, 15q11.2 and 15q13.3 show less clustering in the Icelandic population than would be expected if they were selectively neutral (Supplementary Information). All these CNVs are flanked by large and complex LCR sequences (Supplementary Figs 1, 3 and 4). The LCR can mediate non-allelic homologous recombination, which may result in loss or gain of genomic segments. Through this process CNVs under negative selection can be maintained at low frequency in the population. Other mechanisms for generating rearrangements[26] cannot be excluded. For none of the deletions associated with schizophrenia are we able to pinpoint which LCRs are mediating the non-allelic homologous recombination owing to the complexity of the regions flanking the deletions. Notably, the same CNVs are implicated in schizophrenia and autism and an important area for future study is to determine whether deletions conferring schizophrenia-like syndrome should be considered as classical schizophrenia or new microdeletion syndromes.

In the present study we searched for variants that we think are most likely to confer risk of schizophrenia, namely large recurrent CNVs likely to be under negative selection pressure, rather than testing a large number of selectively neutral CNVs. It is important to identify all recurrent CNVs under negative selection and test those variants for enrichment in well powered samples of schizophrenia cases as well as cases of autism and mental retardation. To determine diagnostic and treatment implications it is also important to study the CNVs conferring risk with respect to drug response, disease progression and symptomatology. Two of the three deletions described here confer high risk of schizophrenia (OR > 11), whereas the third is more common and with more modest risk (OR $= 2.73$). Already identified CNVs associating with schizophrenia may point the way towards underlying pathogenic pathways in the disease; furthermore, high-resolution scans for copy number variants may well identify more CNVs associated with the disease, and given the high odds ratio, these are likely to be clinically useful in diagnosis and risk assessment. Although the CNVs reported here only account for a very small fraction of the genetic risk of schizophrenia, this is an exciting step towards what promises to be a fruitful field for further investigation.

Note added in proof: Samples from the University of Aberdeen were genotyped independently by the International Schizophrenia Consortium[28].

METHODS SUMMARY

Subjects. This study was approved by the National Bioethics Committees or the Local Research Ethical Committees and Data Protection Commissions or laws in the respective countries, Iceland, United Kingdom (Scotland and England), Germany, Finland, Italy, Denmark, Norway, The Netherlands and China. Informed consent was obtained from all patients (Supplementary Information). Of the 4,718 genotyped cases, 4,571 were diagnosed with schizophrenia, 39 with unspecified functional psychosis, 86 with schizoaffective disorder, 10 with schizophreniform and 12 with persistent delusional disorders (Supplementary Information).

Genotyping. The SGENE samples (samples from six European groups, http://www.sgene.eu/) typed on the HumanHap300 chip were used in phase I of the study (Table 1). In phase II, (Table 2) CNV data were derived from the HumanHap300 chip, the HumanHap550 chip, the Affymetrix GeneChip(r) GenomeWide SNP 6.0 or dosage measured using Taqman probes[27]. The Scottish samples in Table 2 were typed at Duke University (HumanHap550) in collaboration with GlaxoSmithKline as were 420 of the German samples, all from Munich (HumanHap300). The remaining CNV data (HumanHap550) from Germany (Table 2, $n = 491$) were obtained from the University of Bonn. Norwegian samples (Affymetrix GeneChip(r) GenomeWide SNP 6.0 array) were analysed using the Affymetrix Power Tools 1.8.0. Dosage data for Danish and Chinese samples were generated at deCODE using Taqman assays[27]. Samples with CNVs were verified by genotyping respective samples using the HumanCNV370 chip.

Statistical analysis. For the genome-wide study of *de novo* CNV associating with schizophrenia the significance threshold was set at 7.6×10^{-4}, which is approximately 0.05/66, the number of *de novo* CNVs identified and tested. All *P*-values are two-sided and there is no overlap between samples in Tables 1 and 2. An exact conditional Cochran–Mantel–Haenszel test (conditional on the strata margins) was used to test for association of schizophrenia and the various CNVs.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 17 April; accepted 8 July 2008.
Published online 30 July 2008; corrected 11 September 2008 (details online).

1. Vogel, H. P. Fertility and sibship size in a psychiatric patient population. A comparison with national census data. *Acta Psychiatr. Scand.* **60**, 483–503 (1979).
2. Weiss, L. A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358**, 667–675 (2008).
3. Shifman, S. *et al.* Genome-wide association identifies a common variant in the reelin gene that increases the risk of schizophrenia only in women. *PLoS Genet.* **4**, e28 (2008).
4. Lu, X. *et al.* Clinical implementation of chromosomal microarray analysis: summary of 2513 postnatal cases. *PLoS ONE* **2**, e327 (2007).
5. de Vries, B. B. *et al.* Diagnostic genome profiling in mental retardation. *Am. J. Hum. Genet.* **77**, 606–616 (2005).
6. Millar, J. K. *et al.* Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Hum. Mol. Genet.* **9**, 1415–1423 (2000).
7. Sharp, A. J. *et al.* Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nature Genet.* **38**, 1038–1042 (2006).
8. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539–543 (2008).
9. Ni, X. *et al.* Connexin 50 gene on human chromosome 1q21 is associated with schizophrenia in matched case control and family-based studies. *J. Med. Genet.* **44**, 532–536 (2007).
10. Brzustowicz, L. M., Hodgkinson, K. A., Chow, E. W., Honer, W. G. & Bassett, A. S. Location of a major susceptibility locus for familial schizophrenia on chromosome 1q21-q22. *Science* **288**, 678–682 (2000).
11. Gurling, H. M. *et al.* Genomewide genetic linkage analysis confirms the presence of susceptibility loci for schizophrenia, on chromosomes 1q32.2, 5q33.2, and 8p21–22 and provides support for linkage to schizophrenia, on chromosomes 11q23.3–24 and 20q12.1–11.23. *Am. J. Hum. Genet.* **68**, 661–673 (2001).

12. Hwu, H. G., Liu, C. M., Fann, C. S., Ou-Yang, W. C. & Lee, S. F. Linkage of schizophrenia with chromosome 1q loci in Taiwanese families. *Mol. Psychiatry* **8,** 445–452 (2003).
13. Zheng, Y. *et al.* A two-stage linkage analysis of Chinese schizophrenia pedigrees in 10 target chromosomes. *Biochem. Biophys. Res. Commun.* **342,** 1049–1057 (2006).
14. Murthy, S. K. *et al.* Detection of a novel familial deletion of four genes between BP1 and BP2 of the Prader-Willi/Angelman syndrome critical region by oligo-array CGH in a child with neurological disorder and speech impairment. *Cytogenet. Genome Res.* **116,** 135–140 (2007).
15. Rogers, S. J., Wehner, D. E. & Hagerman, R. The behavioral phenotype in fragile X: symptoms of autism in very young children with fragile X syndrome, idiopathic autism, and other developmental disorders. *J. Dev. Behav. Pediatr.* **22,** 409–417 (2001).
16. Dimitropoulos, A. & Schultz, R. T. Autistic-like symptomatology in Prader-Willi syndrome: a review of recent findings. *Curr. Psychiatry Rep.* **9,** 159–164 (2007).
17. Nowicki, S. T. *et al.* The Prader-Willi phenotype of fragile X syndrome. *J. Dev. Behav. Pediatr.* **28,** 133–138 (2007).
18. Borghgraef, M., Fryns, J. P. & van den Berghe, H. The female and the fragile X syndrome: data on clinical and psychological findings in 7 fra(X) carriers. *Clin. Genet.* **37,** 341–346 (1990).
19. Thompson, N. M. *et al.* Neurobehavioral characteristics of CGG amplification status in fragile X females. *Am. J. Med. Genet.* **54,** 378–383 (1994).
20. Hancock, M. L., Canetta, S. E., Role, L. W. & Talmage, D. A. Presynaptic type III neuregulin1-ErbB signaling targets α7 nicotinic acetylcholine receptors to axons. *J. Cell Biol.* **181,** 511–521 (2008).
21. Freedman, R. *et al.* Linkage of a neurophysiological deficit in schizophrenia to a chromosome 15 locus. *Proc. Natl Acad. Sci. USA* **94,** 587–592 (1997).
22. Erdogan, F. *et al.* Characterization of a 5.3 Mb deletion in 15q14 by comparative genomic hybridization using a whole genome "tiling path" BAC array in a girl with heart defect, cleft palate, and developmental delay. *Am. J. Med. Genet. A* **143,** 172–178 (2007).
23. Fernandes, C., Hoyle, E., Dempster, E., Schalkwyk, L. C. & Collier, D. A. Performance deficit of α7 nicotinic receptor knockout mice in a delayed matching-to-place task suggests a mild impairment of working/episodic-like memory. *Genes Brain Behav.* **5,** 433–440 (2006).
24. Karayiorgou, M. *et al.* Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11. *Proc. Natl Acad. Sci. USA* **92,** 7612–7616 (1995).
25. Murphy, K. C. Schizophrenia and velo-cardio-facial syndrome. *Lancet* **359,** 426–430 (2002).
26. Lee, J. A., Carvalho, C. M. & Lupski, J. R. A. DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131,** 1235–1247 (2007).
27. Bieche, I. *et al.* Novel approach to quantitative polymerase chain reaction using real-time detection: application to the detection of gene amplification in breast cancer. *Int. J. Cancer* **78,** 661–666 (1998).
28. The International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* doi:10.1038/nature07239 (this issue).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We want to thank the subjects and their relatives and staff at the recruitment centres. This work was sponsored by EU grant LSHM-CT-2006-037761 (Project SGENE), Simons Foundation and R01MH71425-01A1. Genotyping of the Dutch samples was sponsored by NIMH funding, R01 MH078075. This work was also supported by the Chinese National Natural Science Foundation and the National Genomic Network (NGFN-2) of the German Federal Ministry of Education and Research (BMBF). M.M.N. received support from the Alfried Krupp von Bohlen und Halbach-Stiftung. We are grateful to S. Schreiber and M. Krawczak for providing genotype data for PopGen controls, and to K.-H. Jöckel and R. Erbel for providing control individuals from the Heinz Nixdorf Recall Study. We thank L. Priebe and M. Alblas for technical assistance and analysis of CNV data from Bonn.

Author Contributions H.S., D.R., E.S., D.C., L.P., D.S.C. and K.S. wrote the first draft of the paper. M.H., B.B.M., K.D.J., P.M., I.G., H.-J.M., A.H., A.C.N., C.C., G.F., N.W., J.L., J.S., A.T., T.T., E.B., M.D.F., R.M., M.R, S.T., M.W., T.L., C.V., T.W.M., A.G.W., H.U., S.D., I.M., J.O., O.A.A., A.G., M.R., R.O., J.B., R.S.K., L.A.K. and B.F. recruited, diagnosed and gathered phenotypes. H.S., D.R., R.F., E.S., T.S., C.F., P.M., T.T, J.R.G., U.T., H.P., D.G., T.W., D.C., L.P., D.S.C. and K.S. planned, supervised and coordinated the work. H.S., S.C., A.I., S.S., A.G., T.E.T., O.P.H.P., B.V.H., D.G., K.V.S., M.M.N., T.H. and A.K. analysed the data. A.S., A.J., A.J., A.B., S.M. and T.B. performed genotyping and experimental work. All authors contributed to the current version of the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.S.C. (d.stclair@abdn.ac.uk) or K.S. (kari.stefansson@decode.is).

Genetic Risk and Outcome in Psychosis (GROUP)

René S. Kahn[1], Don H. Linszen[2], Jim van Os[3], Durk Wiersma[4], Richard Bruggeman[4], Wiepke Cahn[1], Lieuwe de Haan[2], Lydia Krabbendam[3] & Inez Myin-Germeys[3]

[1]Department of Psychiatry, Rudolf Magnus Institute of Neuroscience, University Medical Centre Utrecht, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands. [2]Academic Medical Centre University of Amsterdam, Department of Psychiatry, Amsterdam, The Netherlands. [3]Maastricht University Medical Centre, South Limburg Mental Health Research and Teaching Network, Maastricht, The Netherlands. [4]University Medical Centre Groningen, Department of Psychiatry, University of Groningen, The Netherlands.

METHODS

***De novo* CNV analysis.** To uncover *de novo* CNVs genome-wide we analysed data from a population-based sample of 2,160 trios and 5,558 parent–offspring pairs, totalling 9,878 transmissions. Samples were genotyped using the Illumina HumanHap300 or the HumanCNV370 chips. To identify *de novo* deletions, we combined two complementary methods: DosageMiner, a Hidden Markov Model algorithm based on intensity data that is similar to that reported previously[29], and a procedure using inheritance errors and the neighbouring genotype configurations comparable to that described previously[30]. When only one parent was typed, using genotype information allowed us to identify deletions as putatively *de novo* by assessment of regional parental heterozygosity. To identify *de novo* duplications we analysed CNV data from the 2,160 trios using DosageMiner.

CNVs in phase 1 were identified by using DosageMiner, software developed by deCODE genetics, and loss of heterozygosity analysis. CNV events stand out in the data from two perspectives. First, all sample intensities for SNPs/probes within a CNV should be increased or decreased relative to neighbouring SNPs/probes that are not in a CNV region; second, CNVs can be detected from the transmission from parent to child. To determine deviations in signal intensity we start by normalizing the intensities. The normalized intensities for each colour channel were determined by a fit of the following equation: $\log(x_{ij}) = f(\alpha_i, gc(j)) + \mu_{j,gen(i,j)} + \beta_i + \varepsilon_{ij}$, where i is sample index, j is SNP index, x_{ij} is colour intensity for sample i in SNP j, $gc(j)$ is an indicator of G+C content around SNP j, f is a smooth function of G+C content, α_i are sample-specific parameters for G+C content, $gen(i,j)$ is the genotype for sample i for SNP j, $\mu_{j,gt}$ is the SNP effect for genotype gt and SNP j, β_i is sample effect, and ε_{ij} is the unexplained part of the signal, including noise. The same model with another set of parameters is used for the other colour y_{ij}. A generalized additive model[31] is used to fit the smooth function f. After fitting the model, the data are normalized by removing the systematic model components. We consider a region to be a deletion/duplication if the average intensity over at least ten markers in a region falls below/above an empirically determined threshold.

To identify regions demonstrating loss of heterozygosity (LOH), markers are split into three classes: (1) shows LOH; (2) inconsistent with LOH; and (3) consistent with LOH. Class 3 is further split into two subclasses: (a) consistent with transmitted LOH; (b) consistent with *de novo* LOH. A marker shows LOH if a child is homozygous for one allele and a parent is homozygous for the other allele. A marker is inconsistent with LOH if the child is heterozygous. A marker is consistent with LOH if the child is homozygous and the parent is homozygous for the same allele or heterozygous. In case the parent is homozygous for the same allele as the child, the marker is consistent with transmitted LOH, and in case the parent is homozygous for the other allele, the marker is only consistent with *de novo* LOH.

A stretch containing a single marker showing LOH is likely to be due to a genotyping error, but because our genotyping error rate is low and independent of position on the genome, the occurrence of more than one marker showing LOH in a consecutive stretch on the genome is more likely to be evidence of a deletion in the child. We consider a region to be a putative deletion if at least two markers are showing LOH and *de novo* if consistent with *de novo* LOH.

We analysed 9,878 offspring–parent pairs consisting of a total of 7,718 offspring and 7,121 parents. Using LOH analysis we define a candidate deleted region if more than one marker shows inheritance error within a region of homozygous markers. We identified a total of 270 candidate *de novo* deletions using this approach. Of these, 80 belong to six distinct individuals which all had multiple regions identified as *de novo* deletions on the same chromosome. On further inspection of the data for these individuals we concluded that they were examples of uniparental disomy. Once these individuals were removed, the remaining 190 putative *de novo* deletions were compared with the output of DosageMiner, and 55 were consistently called deletions by both approaches. These 55 *de novo* deletions represent 51 loci (Supplementary Table 3). In addition 15 large duplications, of 20 or more consecutive markers, were also identified in the trio sample by DosageMiner (Supplementary Table 3).

Dosage measurements using Taqman assays. The Danish and Chinese samples in Table 2 were typed using Taqman assays[27]. The 1q21.1 assay (*PRK* assay) and 15q11.2 assay (*NIPA2* assay) were designed using Primer Express software. Applied Biosystems provided FAM-labelled probes for the assay which were run as described previously[27]. For the reference assay we used a probe in the *CFTR* gene and used the same protocol. The second reference assay, RNASEP ready to use assay, was supplied by Applied Biosystems. Samples identified with deletions or duplications by the Taqman dosage measurements were confirmed by typing the sample on the Illumina HumanCNV370 array.

Probe and primers used for the 1q21.1 assay: 6FAM-CCTGCTGTGTGGGCT-MGB (minor groove binder), PRK-F, CCTTCAGACCAGCGGATAACA and PRK-R, CATGGCAGCAGGATTTGGA. Probe and primers used for the 15q11.2 assay: 6FAM-CAGAGCAGATTGTTATGTAC-MGB, NIPA2-F, GACT GAAAACGCGCCGATT and NIPA2-R, CCATGGACAGACAAACATTCTTG. Probe and primers used for the CFTR assay: 6FAM-ATTAAGCACAGTG GAAGAA-MGBNFQ (minor groove binder non-fluorescent quencher), CFTR-F, AACTGGAGCCTTCAGAGGGTAA and CFTR-R, CCAGGAAAACT GAGAACAGAATGA.

Plates were sealed with optical adhesive cover (Applied Biosystems) and the real-time PCR carried out on an ABI 7900 HT machine for 40 cycles of 15 s at 95 °C and 1 min at 60 °C starting out with an initial step of 10 min at 95 °C.

29. Colella, S. *et al.* QuantiSNP: an objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.* **35,** 2013–2025 (2007).
30. Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E. & Pritchard, J. K. A high-resolution survey of deletion polymorphism in the human genome. *Nature Genet.* **38,** 75–81 (2006).
31. Hastie, T. & Tibshirani, R. Generalized additive models. *Stat. Sci.* **1,** 297–310 (1986).

LETTERS

Rare chromosomal deletions and duplications increase risk of schizophrenia

The International Schizophrenia Consortium*

Schizophrenia is a severe mental disorder marked by hallucinations, delusions, cognitive deficits and apathy, with a heritability estimated at 73–90% (ref. 1). Inheritance patterns are complex, and the number and type of genetic variants involved are not understood. Copy number variants (CNVs) have been identified in individual patients with schizophrenia[2–7] and also in neurodevelopmental disorders[8–11], but large-scale genome-wide surveys have not been performed. Here we report a genome-wide survey of rare CNVs in 3,391 patients with schizophrenia and 3,181 ancestrally matched controls, using high-density microarrays. For CNVs that were observed in less than 1% of the sample and were more than 100 kilobases in length, the total burden is increased 1.15-fold in patients with schizophrenia in comparison with controls. This effect was more pronounced for rarer, single-occurrence CNVs and for those that involved genes as opposed to those that did not. As expected, deletions were found within the region critical for velo-cardio-facial syndrome, which includes psychotic symptoms in 30% of patients[12]. Associations with schizophrenia were also found for large deletions on chromosome 15q13.3 and 1q21.1. These associations have not previously been reported, and they remained significant after genome-wide correction. Our results provide strong support for a model of schizophrenia pathogenesis that includes the effects of multiple rare structural variants, both genome-wide and at specific loci.

The International Schizophrenia Consortium was established to promote rapid progress towards the identification of genetic causes underlying schizophrenia. The consortium is composed of investigators from the University of Aberdeen, Cardiff University, the University of Edinburgh, Karolinska Institutet, Massachusetts General Hospital, the University of North Carolina-Chapel Hill, the Queensland Institute of Medical Research, the University of Southern California, the Stanley Center for Psychiatric Research at the Broad Institute of Harvard and MIT, Trinity College Dublin and University College London.

We surveyed single nucleotide polymorphisms (SNPs) and CNVs using the Affymetrix Genome-Wide Human SNP 5.0 and 6.0 arrays in European cases of schizophrenia and in ancestrally matched controls (Table 1 and Supplementary Information)[13]. On the basis of the genome-wide SNP data there was no evidence of major population stratification within each site[14] (data not shown). Intensity data from both SNP and CNV probes were used to identify autosomal deletions and duplications, based on a hidden Markov model[15].

This study focused on rare but highly penetrant structural variation in schizophrenia, following a natural extension of the classical medical genetic approach. Common CNVs are better identified with different algorithms and are better tested for association separately[13,15]. Considering CNVs that were present in less than 1% of our total sample, there were 6,753 larger than 100 kilobases (kb) that passed sample and CNV quality filtering (see Supplementary Information and Supplementary Table 1). The median size was 182.1 kb (166.3 kb for deletions, 194.4 kb for duplications), 39% were deletions and the median number per individual was 1. We assessed the impact of rare structural variation on the risk for schizophrenia in two ways: first in terms of an individual's genome-wide burden, and second by searching for specific loci that were significantly associated with disease.

Structural variants have been identified for severe neurodevelopmental disorders[9–11,16,17]. Because it has been postulated that schizophrenia might, at least in part, have a developmental aetiology[18], we posited a role for CNVs in schizophrenia, as have others[2–6]. Several loci have been identified, including variants containing genes with neurodevelopmental roles[2–5]. However, a critical question is the extent to which this is a general mechanism for producing schizophrenia in typical clinical populations rather than in cases selected for atypical phenotypic features such as very early onset or mental retardation. This motivated our primary hypothesis: that individuals with schizophrenia have a greater genome-wide burden of CNVs. Considering all CNVs, we observed that cases had a greater average burden than controls (one-sided, empirical $P = 3 \times 10^{-5}$ controlling for array type; Table 2). Controls on average had 0.99 CNVs per person, whereas cases showed a 1.15-fold higher rate.

We next explored this subtle, but highly statistically significant, observation of increased burden. We defined burden in two ways: as the number of CNVs carried by an individual (as above), and also as the number of genes spanned by those CNVs. This second metric (the 'gene count') in fact showed a stronger association with schizophrenia (1.41-fold increase, empirical $P = 2 \times 10^{-6}$) than burden defined simply as the number of CNVs. Characteristics of CNV subgroups studied here are their frequency, type, size, and proximity to a gene (Tables 2 and 3, and Supplementary Table 2). We observed an increased burden across multiple independent subgroups of CNVs,

Table 1 | Study sample characteristics and genotyping platforms

Sample	Ancestry	Case (*n*)	Control (*n*)	Genotyping platform
University of Aberdeen	Scottish	727	694†	5.0
University College London	British	547	n/a‡	5.0
Portuguese Island Collection	Portuguese	333	200†	5.0
Karolinska Institutet	Swedish	622	437	5.0/6.0§
Cardiff University	Bulgarian	479*	646	6.0
Trinity College Dublin	Irish	280	914	6.0
University of Edinburgh	Scottish	403*	290	6.0

Figures are the numbers of cases and controls passing quality control and included in the final analyses. Case samples received a diagnosis of schizophrenia. 'Genotyping platform' indicates Affymetrix array type (5.0 or 6.0).
* Cases were excluded if IQ was less than 70.
† Controls were screened for psychiatric disorders.
‡ University College London control samples genotyped with the Affymetrix 500K two-chip genotyping platform were excluded because CNV data were not available.
§ Swedish cases and controls matched for array type for all analyses.

*Lists of members and affiliations appear at the end of the paper.

Table 2 | Global CNV burden analysis: event type and frequency

			CNV burden (number)			CNV burden (gene count)		
CNV type	Frequency	CNV (*n*)	*P*	Case/control ratio	Baseline rate (controls)	*P*	Case/control ratio	Baseline rate (controls)
Deletions and duplications	All	6,753	3×10^{-5}	1.15	0.99	2×10^{-6}	1.41	2.01
	Single occurrence	890	5×10^{-6}	1.45	0.11	0.0057	1.67	0.32
	2–6 occurrences	2,465	0.0013	1.16	0.35	5×10^{-4}	1.36	0.80
Deletions only	All	2,652	0.11	1.08	0.40	3×10^{-5}	1.55	0.72
	Single occurrence	470	0.011	1.29	0.06	0.005	1.77	0.12
	2–6 occurrences	994	0.048	1.15	0.15	0.13	1.38	0.21
Duplications only	All	4,101	2×10^{-5}	1.20	0.59	10^{-4}	1.28	1.94
	Single occurrence	734	8×10^{-6}	1.58	0.09	0.015	1.60	0.30
	2–6 occurrences	1,532	0.011	1.16	0.22	0.012	1.30	0.69

The table shows an analysis of global CNV burden in cases versus controls. As described in the text, CNVs have previously been filtered for a maximum of about 1% sample frequency. These analyses were further stratified according to type (deletions versus duplications) and frequency (single occurrences and CNVs observed two to six times). Empirical *P* values (one-sided, controlling for array type) are given for two measures of CNV burden (number of CNVs and number of genes affected by CNVs). The average rate in controls (baseline, number of CNVs per person) and the fold increase in cases (case/control ratio) are shown for each analysis. Note that the 'Deletions only' and 'Duplications only' counts are not expected to sum to the 'Deletions and duplications' count for the two lower-frequency groups (see Supplementary Information).

a finding that was more pronounced for rarer CNVs and those involving genes. Deletions and duplications also had somewhat different profiles: the association of deletions varied more noticeably with respect to CNV size and proximity to a gene, whereas duplications showed a more uniform pattern.

A total of 890 CNVs were observed in either a case or a control as a single occurrence. This rarest subset of CNVs would be expected to show enrichment under the model that genetic causes of schizophrenia are individually unique in some proportion of patients. Indeed, this set of CNVs showed a 1.45-fold increase in cases (empirical $P=5\times10^{-6}$). On average, 13.1% of cases of schizophrenia possessed a deletion or duplication observed only once in the sample, in contrast to 10.4% of controls. Under a model in which very rare (occurring in under 1/1,000 individuals) inherited or recurrently *de novo* events increase risk, we would expect to observe a greater overall burden in schizophrenia. Although our study was statistically underpowered to identify the actual loci involved, such variants could in theory be mapped in extremely large samples. In this intermediate group, we observed 2,465 CNVs occurring between two and six times in the total sample, for which there was an increased burden, both for number of CNVs (empirical $P=0.0013$) and gene count (empirical $P=5\times10^{-4}$).

Because several known genomic disorders of the nervous system result from large CNVs, which are often many hundreds of kilobases long[11], we additionally stratified by size of event (Table 3). Of deletions, only larger (more than 500 kb) variants were enriched (empirical $P=3\times10^{-4}$) despite being the least frequent set of CNVs ($n=285$), displaying a 3.57-fold increase in gene count between cases and controls (empirical $P=2\times10^{-5}$). In contrast, shorter duplications showed a stronger association with disease than longer duplications, albeit with a smaller fold increase than deletions (Table 3).

In general, the gene count definition of CNV burden yielded stronger results, particularly for deletions (gene count $P=3\times10^{-5}$ versus number $P=0.11$; Table 2). In fact, dividing all CNVs into two sets, of those that intersect at least one gene and those that do not, we saw an increased burden only in the number of 'genic' CNVs ($P=5\times10^{-6}$; Supplementary Table 2) and not for non-genic CNVs ($P=0.16$). There was a similar trend for CNVs seen two to six times when comparing enrichment in genic and non-genic CNVs ($P=7\times10^{-4}$ and 0.19, respectively) but not single-occurrence CNVs ($P=6\times10^{-4}$ and 6×10^{-4}, respectively). These results may reflect biological distinctions, although they may to some extent also reflect variable performance in CNV detection for different classes of variant.

We conducted a set of analyses to rule out several sources of bias and confounding in the primary genome-wide burden analysis (Supplementary Tables 3–6). Although, in general, low specificity and sensitivity decrease power, of concern here is potential measurement error that varied systematically between cases and controls, leading to spurious results. In this respect, an obvious concern is that both Affymetrix 5.0 and 6.0 arrays were used; as a consequence, we performed all analyses controlling for array type. As described in Supplementary Information, the primary result was also robust to the following. First, in addition to array type, we controlled for sample collection site, genotyping plate and average probe variance. Second, sensitivity analyses showed that no single sample collection site accounted for the observations. Third, we restricted analysis to the most homogeneous 90% of the sample with respect to intra-individual probe variance. Fourth, if differences in CNV burden between cases and controls were purely due to unmeasured confounders, we would not expect an enriched gene count after controlling for the overall extent and rate of CNVs. However, after controlling for overall (genic and non-genic) CNV burden there remained a significantly enriched gene-count burden in patients with schizophrenia.

Our large sample size further enabled us to search for specific CNV regions associated with schizophrenia. One locus previously reported to increase risk for schizophrenia is 22q11.2 (17–21 megabases (Mb)), at which hemizygosity occurs in 1 in every 4,000 live births[12]. These deletions produce a range of clinically heterogeneous phenotypes, including velo-cardio-facial syndrome and DiGeorge syndrome, that

Table 3 | Global CNV burden analysis: event type and size

			CNV burden (number)			CNV burden (gene count)		
CNV type	Size range (kb)	CNV (*n*)	*P*	Case/control ratio	Baseline rate (controls)	*P*	Case/control ratio	Baseline rate (controls)
Deletions and duplications	100 – 200	3,725	0.0017	1.15	0.55	8×10^{-6}	1.35	0.73
	200 – 500	2,156	0.028	1.11	0.32	0.0088	1.25	0.66
	≥500	872	0.0013	1.32	0.12	8×10^{-4}	1.79	0.62
Deletions only	100 – 200	1,612	0.54	1.02	0.25	0.28	1.07	0.26
	200 – 500	755	0.39	1.04	0.12	0.059	1.27	0.14
	≥500	285	3×10^{-4}	1.67	0.03	2×10^{-5}	3.57	0.14
Duplications only	100 – 200	2,113	10^{-4}	1.26	0.30	2×10^{-6}	1.50	0.47
	200 – 500	1,401	0.017	1.14	0.20	0.026	1.24	0.52
	≥500	587	0.11	1.17	0.09	0.17	1.29	0.48

The table shows an analysis of global CNV burden in cases versus controls. CNVs were stratified into three size categories (100–200 kb, 200–500 kb, and 500 kb or more). See Table 2 for further details.

together are known as 22q11.2 deletion syndrome (22q11.2DS)[12]; about 30% of carriers develop psychosis[12]. Previous studies estimated the frequency of 22q11.2 deletions to be 0.6–2.0% in cases of schizophrenia, although many of these studies had technically incomplete characterization of this region[19]. We therefore expected to find examples of 22q11.2 deletions in our sample of 6,572 individuals. The most common form of 22q11.2DS is a 3-Mb loss (about 90% frequency), although a nested 1.5-Mb deletion is also observed (about 7%) along with infrequent (about 3%) atypical deletions[20]. We identified 13 large deletions (more than 500 kb) in cases of schizophrenia within this interval, and none in controls (Supplementary Table 7). Of these, six were consistent with the larger deletion, five were consistent with the shorter deletion, and two were atypical. The 11 samples with typical deletions defined an interval with the strongest association (empirical $P = 0.0017$; genome-wide corrected $P = 0.0046$; odds ratio = 21.6) (Fig. 1a). Controlling for sample collection site or genotyping plate instead of array type did not change the results (Supplementary Table 10). The two other atypical deletions in this region overlap the distal end of the 3-Mb variant. Deletion events within the region were confirmed in all 13 patients by quantitative polymerase chain reaction (qPCR) with three individual assays (Supplementary Fig. 1 and Supplementary Tables 11 and 12). Our findings provide additional evidence that hemizygosity in 22q11.2 is a rare but powerful risk factor for schizophrenia.

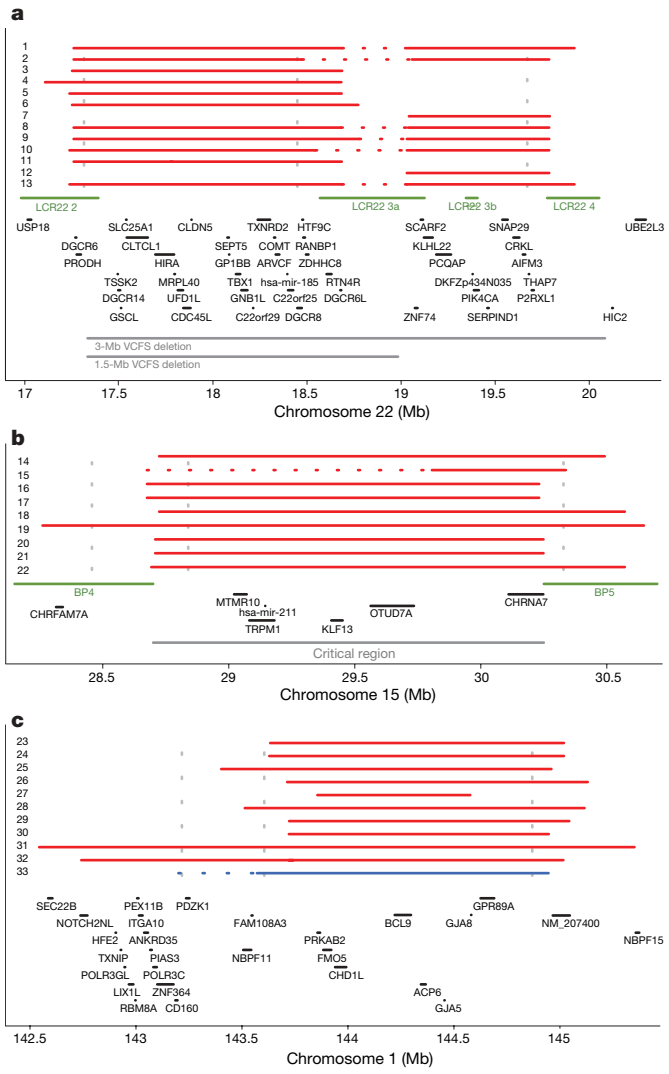


Figure 1 | Regions with excess large deletions in cases. a, The positions of CNVs of more than 500 kb across the chromosome 22q11.2 region. Red lines, case deletions; horizontal dashed sections, qPCR or visual inspection of array intensity data suggest an extended deletion; green lines, locations of low-copy repeats (LCR22-2–LCR22-4).; grey lines, recurrent 3-Mb and 1.5-Mb velocardio-facial syndrome (VCFS) deletions. qPCR primers are marked by vertical dashed lines. **b**, Chromosome 15q13.3 region, as above, except for the locations of breakpoint regions (BP4 and BP5; green) and the critical region defined previously[17] (grey line). **c**, Chromosome 1q21.1, as above, except that a single deletion identified in a control subject is marked by blue line. Genes based on build 35 UCSC browser (Supplementary Table 8).

The larger 22q11.2 deletion harbours 43 genes (Supplementary Table 8). Despite the efforts of many groups, the psychiatric symptoms observed in 22q11.2DS have not been ascribed to a reduced copy number of any individual gene[12]. Variants within *COMT*, which encodes catechol-*O*-methyltransferase, an enzyme responsible for degrading catecholamines, including dopamine, have been implicated in a wide variety of phenotypes, but with inconsistent results[12].

Removing the thirteen 22q11.2DS individuals, we observed a further 271 deletions of more than 500 kb (161 in cases and 110 in controls). Two additional regions (15q13.3 and 1q21.1) were identified that harboured a significant excess of deletions in cases of schizophrenia after correction for multiple testing (Fig. 1b, c; see Supplementary Table 7 for case descriptions). On chromosome 15 (28–31 Mb) there were deletions in nine cases and no controls (empirical $P = 0.0029$; genome-wide corrected $P = 0.046$; odds ratio 17.9). On chromosome 1 (142.5–145.5 Mb) there were ten deletions in cases and one in controls (empirical $P = 0.0076$; genome-wide corrected $P = 0.046$; odds ratio 6.6). All 20 large deletions at 15q13.3 and 1q21.1 were validated by one or more qPCR reactions (Supplementary Fig. 1 and Supplementary Tables 11 and 12). The multiple test correction factors were small as a consequence of our having restricted attention to this small class of rare variants. We did not observe any regions with a corrected $P < 0.05$ for either duplications or smaller (less than 500 kb) deletions. In addition, the primary CNV burden tests remained significant after removing individuals with a deletion at one of these three loci (number $P = 10^{-4}$ and gene count $P = 3 \times 10^{-5}$); for deletions of more than 500 kb specifically, the burden test remained significant for number ($P = 0.02$) but not for gene count ($P = 0.11$).

The large deletions on chromosome 15q13.3 have not previously been associated with schizophrenia. This region does not include the nearby critical region for Prader–Willi/Angelman syndrome[21] but is consistent with the critical region defined by recurrent deletion in cases of mental retardation with seizures that have been reported recently[17]. Furthermore, our estimated breakpoints fall within the segmental duplications reported (BP4 and BP5). In the present study, evidence consistent with mildly impaired cognition was seen in five of the nine patients with deletions, and one individual also had a history of epilepsy (Supplementary Table 7). This broad region has been the focus of previous genetic studies in schizophrenia. The gene *CHRNA7*, encoding the α_7 subunit of the nicotinic acetylcholine receptor, is a candidate based on an initial identification from linkage analysis of auditory evoked potential deficits observed in patients with schizophrenia[22,23].

The deleted region on 1q21.1 is consistent with a previously reported *de novo* deletion in a patient with learning disability and seizures[24] and two patients with autism (one *de novo* and one inherited)[10]. In the present study, three cases had mild cognitive abnormalities and one had a history of epilepsy (Supplementary Table 7). The region contains 27 known genes, most of which are expressed in the brain (Supplementary Table 8), and previous reports have shown linkage[25,26] but there have been no previous reports of CNVs associated with schizophrenia.

Regions of highly homologous segmental duplication flank the deletions we report at 22q11.2, 15q13.3 and 1q21.1. A prominent mechanism for CNV genesis is non-allelic homologous recombination mediated by segmental duplications, resulting in deletions and reciprocal duplications of the interval between segmental duplications[16,27]. Neurodevelopmental and psychiatric syndromes have been associated with deletions and duplications flanked by segmental duplications, many of which occur *de novo*[10,11,17]. Segmental duplications

and non-allelic homologous recombination mediate CNVs at 22q11.2 (ref. 28) and may be involved in the genesis of CNVs at 15q13.3 (ref. 17) and 1q21.1, although other mechanisms may be involved[29].

While this work was under review, Walsh *et al.*[2] reported a higher frequency of cases with CNVs (15%, based on 23 CNVs in 150 patients with schizophrenia) than in controls (5%). Of the 21 autosomal case CNVs identified in that report, we observed overlapping control CNVs at six loci (for example *DLG2* and *PTPRM*; Supplementary Table 9), illustrating that large sample numbers are needed to conclude that any one particular CNV or implicated gene can cause schizophrenia. Our global burden analysis demonstrated that, on aggregate, single-occurrence and very rare (under about 1 in 1,000) CNVs have increased rates in cases of schizophrenia in comparison with controls, in line with Walsh *et al.*[2]. This indicates that at least some of these rare CNVs seen in cases but not in controls are probably risk factors for schizophrenia, although like Walsh *et al.* we are unable to identify which. Some examples of possible risk CNVs that were observed multiple times only in cases include deletions at 12p11.23 (four cases) and 16p12.2–12.1 (four cases). These deletions were more than 500 kb, flanked by segmental duplications and spanning several brain-expressed genes. In addition, duplications in two genes relevant to neural development and growth (*NOTCH1* and p21-activated kinase 7, *PAK7*) were found in five and six cases, respectively, and no controls. Furthermore, we identified CNVs at two recently reported loci, *NRXN1* and *CNTNAP2* (refs 3, 5) (Supplementary Fig. 2).

The aetiology of schizophrenia has been vigorously debated. We now have strong and replicated[2] evidence that individuals with schizophrenia have a greater burden of structural variation across their genomes. Our data show that CNVs in at least three loci act as strong risk factors for schizophrenia in a minority of individuals. We can therefore now posit that some cases of schizophrenia are 'genomic disorders'[16] although we do not yet know whether the risk is specific for schizophrenia as opposed to a more general risk factor for neuropsychiatric or central nervous system illness.

Exactly how a subtle, 1.15-fold increase in CNV burden translates mechanistically into illness in a given patient is currently unknown. We also do not know whether common genetic variants of more subtle effect are components of the aetiology of schizophrenia, an empirical question that we and others are addressing. Similarly, we do not know how environmental risk or protective factors might act in concert with specific CNVs or with the overall burden of CNVs.

A critically important goal will be to determine the full clinical and phenotypic spectrum in carriers of these deletions. Our data provide preliminary evidence of a variable phenotype in patients with schizophrenia who would otherwise be regarded as clinically typical. Examining the role of these variants in related psychotic disorders, such as bipolar disorder, is imperative. Further work explicating the epidemiology and mechanism of these variants in schizophrenia may ultimately lead to a role for them in genetic counselling and understanding disease biology.

Note added in proof: Samples from the University of Aberdeen were genotyped independently by the SGENE Consortium[30].

METHODS SUMMARY

Cases satisfied DSM-IV[31] or ICD-10 (ref. 32) criteria for schizophrenia and were broadly representative of clinical cases in contact with psychiatric services. DNA was extracted from whole blood, with approval from institutional review boards. CNVs were identified with the Birdseye package[15] and analysed with PLINK v1.03 (ref. 14). See Supplementary Information for details. A list of all CNVs passing quality control is available (http://pngu.mgh.harvard.edu/isc/).

Received 14 May; accepted 8 July 2008.
Published online 30 July 2008.

1. Sullivan, P. F., Kendler, K. S. & Neale, M. C. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch. Gen. Psychiatry* **60,** 1187–1192 (2003).
2. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320,** 539–543 (2008).
3. Kirov, G. *et al.* Comparative genome hybridization suggests a role for NRXN1 and APBA2 in schizophrenia. *Hum. Mol. Genet.* **17,** 458–465 (2008).
4. Flomen, R. H. *et al.* Association study of CHRFAM7A copy number and 2 bp deletion polymorphisms with schizophrenia and bipolar affective disorder. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **141,** 571–575 (2006).
5. Friedman, J. I. *et al.* CNTNAP2 gene dosage variation is associated with schizophrenia and epilepsy. *Mol. Psychiatry* **13,** 261–266 (2008).
6. Wilson, G. M. *et al.* DNA copy-number analysis in bipolar disorder and schizophrenia reveals aberrations in genes involved in glutamate signaling. *Hum. Mol. Genet.* **15,** 743–749 (2006).
7. Moon, H. J. *et al.* Identification of DNA copy-number aberrations by array-comparative genomic hybridization in patients with schizophrenia. *Biochem. Biophys. Res. Commun.* **344,** 531–539 (2006).
8. Marshall, C. R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82,** 477–488 (2008).
9. Sebat, J. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316,** 445–449 (2007).
10. Weiss, L. A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358,** 667–675 (2008).
11. Lee, J. A. & Lupski, J. R. Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron* **52,** 103–121 (2006).
12. Williams, N. M., O'Donovan, M. C. & Owen, M. J. Chromosome 22 deletion syndrome and schizophrenia. *Int. Rev. Neurobiol.* **73,** 1–27 (2006).
13. McCarroll, S. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genet.*. (in the press).
14. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
15. Korn, J. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms, and rare CNVs. *Nature Genet.*. (in the press).
16. Lupski, J. R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14,** 417–422 (1998).
17. Sharp, A. J. *et al.* A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nature Genet.* **40,** 322–328 (2008).
18. Ross, C. A. & Pearlson, G. D. Schizophrenia, the heteromodal association neocortex and development: potential for a neurogenetic approach. *Trends Neurosci.* **19,** 171–176 (1996).
19. Karayiorgou, M. *et al.* Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11. *Proc. Natl Acad. Sci. USA* **92,** 7612–7616 (1995).
20. Shaikh, T. H. *et al.* Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Hum. Mol. Genet.* **9,** 489–501 (2000).
21. Butler, M. G., Fischer, W., Kibiryeva, N. & Bittel, D. C. Array comparative genomic hybridization (aCGH) analysis in Prader–Willi syndrome. *Am. J. Med. Genet. A* **146,** 854–860 (2008).
22. Freedman, R. *et al.* Linkage of a neurophysiological deficit in schizophrenia to a chromosome 15 locus. *Proc. Natl Acad. Sci. USA* **94,** 587–592 (1997).
23. Xu, J. *et al.* Evidence for linkage disequilibrium between the α_7-nicotinic receptor gene (*CHRNA7*) locus and schizophrenia in Azorean families. *Am. J. Med. Genet.* **105,** 669–674 (2001).
24. Sharp, A. J. *et al.* Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nature Genet.* **38,** 1038–1042 (2006).
25. Brzustowicz, L. M., Hodgkinson, K. A., Chow, E. W., Honer, W. G. & Bassett, A. S. Location of a major susceptibility locus for familial schizophrenia on chromosome 1q21–q22. *Science* **288,** 678–682 (2000).
26. Gurling, H. M. *et al.* Genomewide genetic linkage analysis confirms the presence of susceptibility loci for schizophrenia, on chromosomes 1q32.2, 5q33.2, and 8p21–22 and provides support for linkage to schizophrenia, on chromosomes 11q23.3–24 and 20q12.1–11.23. *Am. J. Hum. Genet.* **68,** 661–673 (2001).
27. Stankiewicz, P. & Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18,** 74–82 (2002).
28. Shaikh, T. H. *et al.* Low copy repeats mediate distal chromosome 22q11.2 deletions: sequence analysis predicts breakpoint mechanisms. *Genome Res.* **17,** 482–491 (2007).
29. Lee, J. A., Carvalho, C. M. & Lupski, J. R. A. DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131,** 1235–1247 (2007).
30. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* doi: 10.1038/nature07229 (this issue).
31. DSM-IV. *Diagnostic and Statistical Manual of Mental Disorders* 4th edn (American Psychiatric Association, 2000).
32. ICD-10. *International Statistical Classification of Diseases and Related Health Problems* 10th revision (World Health Organization, 2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the patients and families who contributed their time and DNA to these studies, and also D. Altshuler and members of the Medical and Population Genetics group at the Broad Institute of Harvard and Massachusetts Institute of Technology for valuable discussion. The group at the Stanley Center for

Psychiatric Research at the Broad Institute was supported by the Stanley Medical Research Institute (E.M.S.), the Sylvan C. Herman Foundation (E.M.S.), and MH071681 (P.S.). The Cardiff University group was supported by a Medical Research Council (UK) Programme grant and the National Institutes of Mental Health (USA) (CONTE: 2 P50 MH066392-05A1). The group at Karolinska Institutet was supported by the Swedish Council for Working Life and Social Research (FO 184/2000; 2001-2368). The Massachusetts General Hospital group was supported by the Stanley Medical Research Institute (P.S.), MH071681 (P.S.) and a Narsad Young Investigator Award (S.P.). The group at the Queensland Institute of Medical Research was supported by the Australian National Health and Medical Research Council. The Trinity College Dublin group was supported by Science Foundation Ireland, the Health Research Board (Ireland), the Stanley Medical Research Institute and the Wellcome Trust; Irish controls were supplied by J. McPartlin from the Trinity College Biobank. The work at the University of Aberdeen was partly funded by GlaxoSmithKline and Generation Scotland, Genetics Health Initiative. The University College London clinical and control samples were collected with support from the Neuroscience Research Charitable Trust, the Camden and Islington Mental Health and Social Care Trust, East London and City Mental Heath Trust, the West Berkshire NHS Trust, the West London Mental Health Trust, Oxfordshire and Buckinghamshire Mental Health Partnership NHS Trust, South Essex Partnership NHS Foundation Trust, Gloucestershire Partnership NHS Foundation Trust, Mersey Care NHS Trust, Hampshire Partnership NHS Trust and the North East London Mental Health Trust. The collection of the University of Edinburgh cohort was supported by grants from the Wellcome Trust, London, and the Chief Scientist Office of the Scottish Executive. The group at the University of North Carolina, Chapel Hill, was supported by MH074027, MH077139 and MH080403, the Sylvan C. Herman Foundation (P.F.S.) and the Stanley Medical Research Institute (P.F.S.) The group at the University of Southern California thanks the patients and their families for their collaboration, and acknowledges the support of the National Institutes of Mental Health and the Department of Veterans Affairs.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.S. (sklar@chgr.mgh.harvard.edu).

The International Schizophrenia Consortium

Manuscript preparation Jennifer L. Stone[1–4], Michael C. O'Donovan[5], Hugh Gurling[6], George K. Kirov[5], Douglas H. R. Blackwood[7], Aiden Corvin[8], Nick J. Craddock[5], Michael Gill[8], Christina M. Hultman[9,10], Paul Lichtenstein[9], Andrew McQuillin[6], Carlos N. Pato[11], Douglas M. Ruderfer[1–4], Michael J. Owen[5], David St Clair[12] , Patrick F. Sullivan[13], Pamela Sklar[1–4], Shaun M. Purcell[1–4] (Leader); **Data analysis** Jennifer L. Stone[1–4], Douglas M. Ruderfer[1–4], Joshua Korn[3,4], George K. Kirov[5], Stuart Macgregor[14], Andrew McQuillin[6], Derek W. Morris[8], Colm T. O'Dushlaine[8], Mark J. Daly[2–4], Peter M. Visscher[14], Peter A. Holmans[5], Michael C. O'Donovan[5], Patrick F. Sullivan[13], Pamela Sklar[1–4], Shaun M. Purcell[1–4] (Leader); **Management committee** Hugh Gurling[6], Aiden Corvin[8], Douglas H. R. Blackwood[7], Nick J. Craddock[5], Michael Gill[8], Christina M. Hultman[9,10], George K. Kirov[5], Paul Lichtenstein[9], Andrew McQuillin[6], Michael C. O'Donovan[5], Michael J. Owen[5], Carlos N. Pato[11], Shaun M. Purcell[1–4] , Edward M. Scolnick[2,3], David St Clair[12], Jennifer L. Stone[1–4], Patrick F. Sullivan[13], Pamela Sklar[1–4] (Leader); **Cardiff University** Michael C. O'Donovan[5], George K. Kirov[5], Nick J. Craddock[5], Peter A. Holmans[5], Nigel M. Williams[5], Lucy Georgieva[5], Ivan Nikolov[5], N. Norton[5], H. Williams[5], Draga Toncheva[15], Vihra Milanova[16], Michael J. Owen[5]; **Karolinska Institutet/University of North Carolina at Chapel Hill** Christina M. Hultman[9,10], Paul Lichtenstein[9], Emma F. Thelander[9], Patrick Sullivan[13]; **Trinity College Dublin** Derek W. Morris[8], Colm T. O'Dushlaine[8], Elaine Kenny[8], John L. Waddington[17], Michael Gill[8], Aiden Corvin[8]; **University College London** Andrew McQuillin[6], Khalid Choudhury[6], Susmita Datta[6], Jonathan Pimm[6], Srinivasa Thirumalai[18], Vinay Puri[6], Robert Krasucki[6], Jacob Lawrence[6], Digby Quested[19]†, Nicholas Bass[6], David Curtis[20], Hugh Gurling[6]; **University of Aberdeen** Caroline Crombie[21], Gillian Fraser[21], Soh Leh Kwan[12], Nicholas Walker[22], David St Clair[12]; **University of Edinburgh** Douglas H. R. Blackwood[7], Walter J. Muir[7], Kevin A. McGhee[7], Ben Pickard[7], Pat Malloy[7], Alan W. Maclean[7], Margaret Van Beck[7]; **Queensland Institute of Medical Research** Peter M. Visscher[14], Stuart Macgregor[14]; **University of Southern California** Michele T. Pato[11], Helena Medeiros[11], Frank Middleton[23], Celia Carvalho[11], Christopher Morley[23], Ayman Fanous[11,24–26], David Conti[11], James A. Knowles[11], Carlos Paz Ferreira[27], Antonio Macedo[28], M. Helena Azevedo[28], Carlos N. Pato[11]; **Massachusetts General Hospital** Jennifer L. Stone[1–4], Douglas M. Ruderfer[1–4], Joshua Korn[3,4], Steve A. McCarroll[3,4], Mark Daly[2–4], Shaun M. Purcell[1–4], Pamela Sklar[1–4] **Stanley Center for Psychiatric Research and Broad Institute of MIT and Harvard** Shaun M. Purcell[1–4], Jennifer L. Stone[1–4], Kimberly Chambert[2,3], Douglas M. Ruderfer[1–4], Joshua Korn[3,4], Steve A. McCarroll[3,4], Casey Gates[3], Stacey B. Gabriel[3], Scott Mahon[3], Kristen Ardlie[3], Mark J. Daly[2–4], Edward M. Scolnick[2,3], Pamela Sklar[1–4]

[1]Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, 185 Cambridge Street, Boston, Massachusetts 02114, USA. [2]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [3]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. [4]Center for Human Genetic Research, Massachusetts General Hospital, 185 Cambridge Street, Boston, Massachusetts 02114, USA. [5]School of Medicine, Department of Psychological Medicine, School of Medicine, Cardiff University, Cardiff C14 4XN, UK. [6]Molecular Psychiatry Laboratory, Department of Mental Health Sciences, University College London Medical School, Windeyer Institute of Medical Sciences, 46 Cleveland Street, London W1T 4JF, UK. [7]Division of Psychiatry, School of Molecular and Clinical Medicine, University of Edinburgh, Edinburgh EH10 5HF, UK. [8]Neuropsychiatric Genetics Research Group, Department of Psychiatry and Institute of Molecular Medicine, Trinity College Dublin, Dublin 2, Ireland. [9]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, SE-171 77 Stockholm, Sweden. [10]Department of Neuroscience, Psychiatry, Ulleråker, Uppsala University, SE-750 17 Uppsala, Sweden. [11]Center for Genomic Psychiatry, University of Southern California, Los Angeles, California 90033, USA. [12]Institute of Medical Sciences, University of Aberdeen, Foresterhill, Aberdeen AB25 2ZD, UK. [13]Departments of Genetics, Psychiatry, and Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [14]Queensland Institute of Medical Research, 300 Herston Road, Brisbane, Queensland 4006, Australia. [15]Department of Medical Genetics, University Hospital Maichin Dom, Sofia 1431, Bulgaria. [16]Department of Psychiatry, First Psychiatric Clinic, Alexander University Hospital, Sofia 1431, Bulgaria. [17]Molecular and Cellular Therapeutics and RCSI Research Institute, Royal College of Surgeons in Ireland, Dublin 2, Ireland. [18]West Berkshire NHS Trust, 25 Erleigh Road, Reading RG3 5LR, UK. [19]West London Mental Health Trust, Hammersmith and Fulham Mental Health Unit and St Bernard's Hospital, London W6 8RF, UK. [20]Queen Mary College, University of London and East London and City Mental Health Trust, Royal London Hospital, Whitechapel, London E1 1BB, UK. [21]Department of Mental Health, University of Aberdeen, Aberdeen AB25 2ZD, UK. [22]Ravenscraig Hospital, Inverkip Road, Greenock PA16 9HA, UK. [23]State University of New York – Upstate Medical University, Syracuse, New York 13210, USA. [24]Washington VA Medical Center, Washington, DC 20422, USA. [25]Department of Psychiatry, Georgetown University School of Medicine, Washington, DC 20057, USA. [26]Department of Psychiatry, Virginia Commonwealth University, Richmond, VA 23298, USA. [27]Department of Psychiatry, Sao Miguel, 9500-310 Azores, Portugal. [28]Department of Psychiatry, University of Coimbra, 3004-504 Coimbra, Portugal. †Present address: Department of Psychiatry, University of Oxford, Warneford Hospital, Headington, Oxford OX3 7JX, UK.

Vol 455|11 September 2008|doi:10.1038/nature07207

LETTERS

RNA interference screen for human genes associated with West Nile virus infection

Manoj N. Krishnan[1], Aylwin Ng[4], Bindu Sukumaran[1], Felicia D. Gilfoy[5], Pradeep D. Uchil[3], Hameeda Sultana[1], Abraham L. Brass[7], Rachel Adametz[3], Melody Tsui[6], Feng Qian[2], Ruth R. Montgomery[2], Sima Lev[8], Peter W. Mason[5], Raymond A. Koski[9], Stephen J. Elledge[7,10], Ramnik J. Xavier[4]*, Herve Agaisse[3]* & Erol Fikrig[1,10]*

West Nile virus (WNV), and related flaviviruses such as tick-borne encephalitis, Japanese encephalitis, yellow fever and dengue viruses, constitute a significant global human health problem[1]. However, our understanding of the molecular interaction of such flaviviruses with mammalian host cells is limited[1]. WNV encodes only 10 proteins, implying that it may use many cellular proteins for infection[1]. WNV enters the cytoplasm through pH-dependent endocytosis, undergoes cycles of translation and replication, assembles progeny virions in association with endoplasmic reticulum, and exits along the secretory pathway[1–3]. RNA interference (RNAi) presents a powerful forward genetics approach to dissect virus–host cell interactions[4–6]. Here we report the identification of 305 host proteins that affect WNV infection, using a human-genome-wide RNAi screen. Functional clustering of the genes revealed a complex dependence of this virus on host cell physiology, requiring a wide variety of molecules and cellular pathways for successful infection. We further demonstrate a requirement for the ubiquitin ligase CBLL1 in WNV internalization, a post-entry role for the endoplasmic-reticulum-associated degradation pathway in viral infection, and the monocarboxylic acid transporter MCT4 as a viral replication resistance factor. By extending this study to dengue virus, we show that flaviviruses have both overlapping and unique interaction strategies with host cells. This study provides a comprehensive molecular portrait of WNV–human cell interactions that forms a model for understanding single plus-stranded RNA virus infection, and reveals potential antiviral targets.

The host proteins previously reported to facilitate WNV infection (termed host susceptibility factors, HSFs) comprise endosomal transport regulators and vATPase (for entry), eEF1A, TIA-1/TIAR and HMGCR (for replication), and c-Yes (for secretion)[2,3,7–10]. Other host proteins may reduce WNV infection (termed host resistance factors, HRFs): components of the antiviral IRF3 pathway are known HRFs of WNV infection[11]. In this context, we performed a genome-scale small interfering RNA (siRNA)-based screen silencing 21,121 human genes in HeLa cells to comprehensively identify the cellular proteins associated with the early stages of WNV infection, from viral entry through to the intracellular translation of viral RNA. Defects in the later stages of infection, such as replication, assembly or secretion, were not scored by the assay. The assay involved infection of gene-silenced cells with WNV for 24 h, followed by a microscopy-based quantification of the cells immunostained for viral envelope protein to select the candidate host proteins. The screen was done in two steps: a primary screen using a pool of four siRNAs per gene, followed by a validation screen, testing each individual siRNA within the pool separately (for the hits selected in the primary screen) to minimize potential off-target hits (Fig. 1a). The details of the assay and screen are described in Methods and Supplementary Fig. 1.

The RNAi screen identified 283 HSFs and 22 HRFs (of which 273 and 21 respectively are novel; Supplementary Tables 1 and 2). The number of HRFs constituted 7% of the total host factors identified. The identification of (1) some of the known HSFs (vATPase, endosomal transport regulators[3]) and HRFs (IRF3; ref. 11) of WNV infection, and (2) multiple components of macromolecular assemblies—for example, vATPase, the endoplasmic-reticulum-associated degradation (ERAD) pathway, focal adhesion complex (FAC)—validated the reliability of our approach and the *in vitro* model. A cellular map summarizing several screen hits classified into cellular compartments and broad functional association categories is provided in Supplementary Fig. 2.

Of the 283 HSFs, 195 (69%) and 193 (68%) could be classified using biological process and molecular function categories, respectively (Fig. 1b, c; Supplementary Tables 3 and 4). There was a significant enrichment of genes regulating intracellular protein trafficking, cell adhesion and processes associated with the transport of ions and biomolecules. The enriched molecular function categories included hydrolases, transporters, ligases, cell adhesion molecules, membrane traffic proteins and synthases. Among the HSFs, 6 RNA-binding proteins (for example, RBPMS), 20 ubiquitination-related proteins (for example, CBLL1), 21 transcription factors (for example, LDB1), 3 C-type lectins (CLEC7A, CLEC4A and CLEC4C) and 5 protocadherins (for example, PCDHB5) were also present. The RNA-binding protein RBPMS was reported as part of a protein network implicated in Purkinje cell degeneration[12]. Strikingly, the current screen also captured seven other members (COIL, PCP4, UBE2I, LDB1, NUMBL, ATXN7L3 and USP6) interacting with RBPMS (Supplementary Figs 3a, b, and 4a, b).

The screen also identified several genes previously implicated in immunity (Supplementary Tables 1 and 2). Immune related HSFs include β-defensins (DEFB118 and DEFB129, Supplementary Fig. 5a), Rnase L inhibitor ABCE1 (refs 13–15; Supplementary Fig. 5b), LY6E, Zap70, TNFSF13B and DUBA (OTUD5). Among the HRFs, α-defensin DEFA3 and IRF3 are known immune response genes. These findings highlight that defensin family members function as both viral resistance and susceptibility factors[16]. Knockdown of the immunophilin FKBP1B also enhanced WNV infection.

We next determined whether the genes identified from HeLa cells are expressed in tissues targeted by WNV *in vivo* by analysing the

[1]Section of Infectious Diseases, [2]Section of Rheumatology, Department of Internal Medicine, [3]Section for Microbial Pathogenesis, Yale University School of Medicine, New Haven, Connecticutt 06520-8031, USA. [4]Center for Computational and Integrative Biology, and Gastrointestinal Unit, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02114, USA. [5]Department of Pathology, University of Texas Medical Branch, Galveston, Texas 77555, USA. [6]Department of Systems Biology, [7]Department of Genetics, Center for Genetics and Genomics, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. [8]Department of Neurobiology, Weizmann Institute of Science, Israel. [9]L2 Diagnostics, 300 George Street, New Haven, Connecticutt 06511, USA. [10]Howard Hughes Medical Institute, Chevy Chase, Maryland 20815-6789, USA.
*These authors contributed equally to this work.

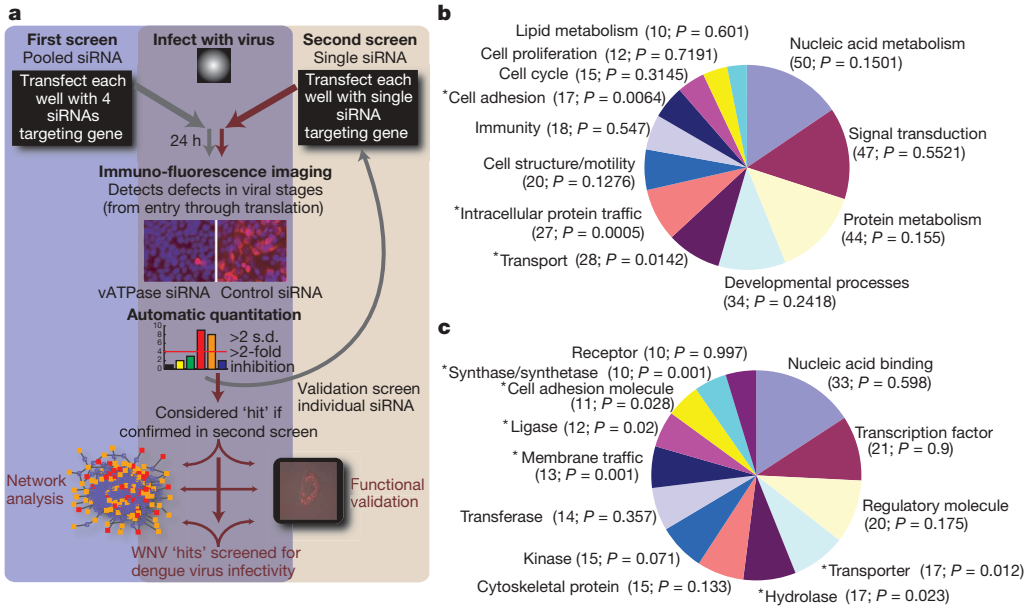


Figure 1 | RNAi screen and bioinformatics. **a**, West Nile virus RNAi screen strategy (see text for description). **b**, **c**, Bioinformatics classification of hits into biological process (**b**) and molecular function (**c**) categories. *Categories found enriched ($P < 0.05$) relative to all the genes examined in the RNAi screen. Only categories with ten or more members are displayed.

expression pattern of the HSFs across 79 tissues (Supplementary Fig. 6). In accordance with the tissue tropism of WNV, 102 (46%) and 64 (29%) HSFs showed enriched expression in immune and central nervous system tissues, respectively (Wilcoxon $P < 0.05$; Supplementary Tables 5 and 6).

Among the 20 ubiquitination-related proteins identified in the screen, the ubiquitin ligase CBLL1 is known to regulate the endocytosis of cell-surface receptors, and therefore we hypothesized that CBLL1 may be involved in the cellular internalization of WNV[17]. *CBLL1* silencing resulted in a marked reduction (82%, $P = 0.05$) of WNV-infected cells (Fig. 2a, b; Supplementary Figs 4a, b, and 7a). In order to test whether CBLL1 is involved in WNV entry, we next examined the kinetics of tetramethyl rhodamine iso-thiocyanate (TRITC)-labelled WNV particle internalization into *CBLL1* silenced cells. Strikingly, there was a ~20-fold ($P < 0.05$) reduction in the number of virus particles present within *CBLL1* silenced cells when analysed from 1 h to 4 h post-incubation (Fig. 2c; Supplementary Fig. 7b). Moreover, virus was seen stuck on the plasma membrane of 17% of *CBLL1* silenced cells (Fig. 2d; Supplementary Figs 7b and 8). As expected, *CBLL1* silencing did not alter WNV replicon translation (Supplementary Fig. 9a). The virus internalization defect of *CBLL1* silenced cells was similar to that observed in cells defective for clathrin dependent endocytosis (CDE), a pathway implicated in WNV entry (Fig. 2d)[2,3]. CDE was ablated by targeting the clathrin adaptor AP3S2, which was also identified in our screen (Supplementary Table 1)[18]. Silencing of the post-entry HSF, vATPase, did not alter the internalization of virus (Fig. 2d). Furthermore, consistent with the involvement of a ubiquitin ligase in WNV entry, depletion of cellular free ubiquitin pool by pretreatment with MG132 (a proteasomal inhibitor) strongly abolished WNV internalization (50-fold, $P = 0.001$) (Fig. 2e; Supplementary Fig. 4b). Notably, proteasome inhibition was also found to interfere with WNV infection at post-internalization steps (Supplementary Fig. 9b). Proteasomal components were also identified in the screen (Supplementary Table 1). Demonstrating WNV specificity, MG132 treatment did not inhibit vesicular stomatitis virus infection (Supplementary Fig. 9c), as reported previously[19]. Collectively, these findings demonstrate that CBLL1 and the proteasome-ubiquitin system are required for the cellular internalization of WNV.

Because the endoplasmic reticulum (ER) is implicated in the intracellular phase of flaviviral life cycle[1], we examined whether WNV co-opts ER components for infection. Network analysis anchoring on ER proteins revealed the presence of several components of the ERAD pathway among the identified HSFs (Fig. 2f). ERAD comprises

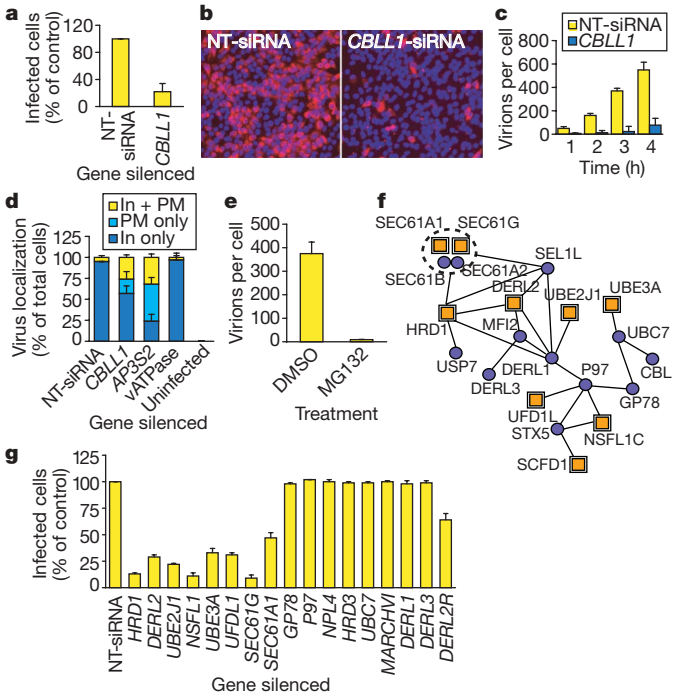


Figure 2 | *CBLL1* and ERAD silencing reduces West Nile virus (WNV) infection. **a–e**, *CBLL1* silenced; **f**, **g**, ERAD silenced. **a**, **b**, *CBLL1* silencing reduces WNV immunostained cells (multiplicity of infection, m.o.i. ≈ 0.3, 10×, Zeiss). Red, virus; blue, nucleus; NT, non-targeting control siRNA. **c**, Time-course analysis showing reduced internalization of TRITC–WNV (m.o.i. ≈ 100) into *CBLL1* silenced cells. **d**, Localization of TRITC–WNV within *CBLL1* silenced cells (percentage of total cells showing the phenotypes, 20 images, ~6 cells each). Plasma membrane alone (PM only), inside the cell (In only), and both PM and inside (In+PM). AP3S2 and vATPase were controls (see text). **e**, MG132 reduces WNV internalization (m.o.i. ≈ 100). **f**, ERAD network. Yellow squares, hits; blue circles, other host proteins within the network neighbourhood. Dotted line, complexes. **g**, ERAD silencing reduces WNV infection (m.o.i. ≈ 0.3). *DERL2R* is an RNAi resistant mutant of *DERL2*. For **a** and **g**, the percentage of infected control NT-siRNA cells (~30%) was set at 100% and used to normalize the percentage infection of siRNA-treated cells (from six fields of ~8,000 cells). Values for **c** and **e** are the number of virus particles per cell (mean of 20 cells). Results are mean ± s.d. from a representative experiment performed in triplicate.

more than ten proteins that retro-transports misfolded proteins from ER to the proteasome[20]. Silencing of several key components or interactors of ERAD (*HRD1*, *DERL2*, *UBE2J1/UBC6*, *UBE3A*, *SEC61G*, *SEC61A1*, *UFD1L* and *NSFL1C*), but not other ERAD components (for example, *DERL1*, *DERL3*, *HRD3*, *NPL4* and *p97*), reduced WNV infected cells up to 89% (Fig. 2g; Supplementary Fig. 4a, b). ERAD was not required for human immunodeficiency virus 2 infection (Supplementary Fig. 10a), highlighting specificity between different viruses. To further validate these results, reduction of viral infection due to silencing of *DERL2* was rescued by transfection with an siRNA-resistant silent mutation-containing variant of DERL2 (Fig. 2g; Supplementary Fig. 10b). We also identified the recently reported ERAD component BCAP31 as an HSF[21]. Functional studies revealed that ERAD is not involved in WNV internalization, endosomal transport[2], or RNA translation; however, there was ~10% reduction in the secretion of progeny virions in ERAD silenced cells (Supplementary Fig. 11a–d, respectively). Interestingly, the simian virus 40 has been shown recently to require the ERAD components DERL1 and SEL1L for uncoating[22]. Together, these results indicate that WNV infection requires a subset of ERAD components at a post-internalization step.

Among the genes whose knockdown enhanced WNV infection, the strongest phenotype was observed when *MCT4* (*SLC16A4*), a plasma membrane transporter of monocarboxylic acids[23], was silenced. Three of the four tested siRNAs targeting *MCT4* resulted in a 10-fold ($P = 0.01$) increase in WNV infected cells at 24 h (Fig. 3a, b; Supplementary Fig. 4a, b). A quantitative PCR-based time-course analysis of the viral genomic RNA (plus-strand) revealed a similar rate of WNV particle internalization into both *MCT4*-repressed and control cells (Fig. 3c). However, replication started at ≤9 h post infection in *MCT4*-silenced cells, whereas in control cells it was delayed until after 12 h (Fig. 3c). Consistent with this, *MCT4* silenced cells (1) had 3, 10, 12 and 18 times ($P < 0.05$) more viral plus-strand RNA at 9 h, 12 h, 15 h and 24 h post infection, respectively (Fig. 3c); (2) immuno-stained for WNV antigens at 9 h (not detectable in control cells until after 12 h; Fig. 3d), and (3) secreted progeny virions by 12 h, whereas control cells did not (Fig. 3e). However, importantly, replication of viral genomic RNA introduced directly to the cytoplasm bypassing the entry stages was not affected by *MCT4* silencing (Supplementary Fig. 12). Collectively, these observations show that the functional activity of *MCT4* delays the temporal transition into the replication phase of endocytosed WNV particles.

We next examined whether the host cell interaction strategies are similar between different members of the genus *Flavivirus* by investigating the effect of silencing all the identified WNV HSFs and HRFs in HeLa cells infected by dengue virus 2 (DENV). We determined that 30 h post-infection of DENV is comparable to the 24 h infection of WNV (Supplementary Fig. 1a, b). Silencing of 36% of the WNV HSFs reduced DENV infection, including previously implicated vATPase and UBE2I (Supplementary Table 1)[3,24]. In contrast, all the 22 WNV HRFs increased DENV infection (Supplementary Table 2). Further supporting pathogen specificity, only five of the host factors impacting WNV infection altered HIV-2 infection (not shown).

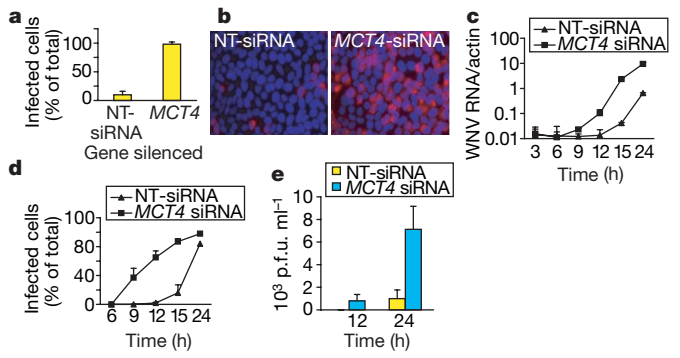


Figure 3 | *MCT4* silencing enhances WNV replication. **a**, **b**, *MCT4* silencing increases the number of WNV immunostained cells (m.o.i. ≈ 0.3, 10×, Zeiss). Red, virus; blue, nucleus. **c**, qPCR of WNV RNA levels in control NT-siRNA and *MCT4* silenced cells (ng viral RNA / ng β-actin). **d**, Immunostaining for WNV E-protein in *MCT4* silenced cells. **e**, WNV secretion from *MCT4* silenced cells, expressed as plaque forming units per ml (p.f.u. ml^{-1}). Values in **a** and **d** are percentage of total cells immunostained for WNV (six images having ~1,000 cells each). Results are mean ± s.d. from a representative experiment performed in triplicate.

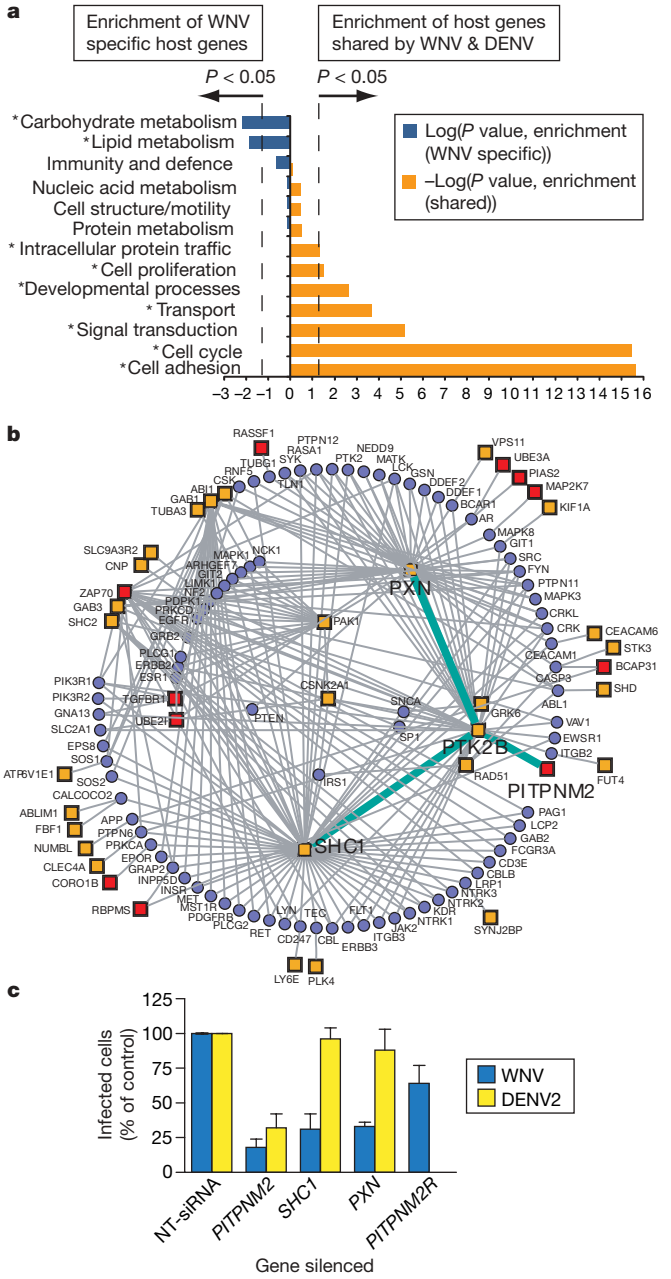


Figure 4 | Interaction of West Nile virus (WNV) and dengue virus (DENV) with host cells. **a**, Classification into biological process categories of HSFs common to both WNV and DENV or specific to WNV. *Categories found enriched ($P < 0.05$) relative to all identified HSFs. **b**, Focal adhesion complex (FAC) network. Cyan line connects the core proteins; yellow squares, WNV specific HSFs; red squares, WNV and DENV shared HSFs; blue circles, other host proteins within the network neighbourhood. **c**, Effect of silencing of *PXN*, *SHC1* and *PITPNM2* on WNV and DENV infection (the percentage of infected control NT-siRNA cells (~30%) was set at 100% and used to normalize the percentage infection of siRNA-treated cells (from six fields of ~8,000 cells). *PITPNM2R* indicates an RNAi-resistant mutant of *PITPNM2*. Results are mean ± s.d. from a representative experiment performed in triplicate.

An assessment of enrichment for biological process categories revealed significant over-representation ($P < 0.05$) of seven key processes in which HSFs are targeted by both WNV and DENV (Fig. 4a), relative to their representation among all HSFs identified. We selected three pathways—ERAD, FAC and histone deacetylase (HDAC)—to compare the conservation between WNV and DENV. There was a near-complete overlap of ERAD component usage shared by both WNV and DENV, with the single exception of HRD1 (Supplementary Fig. 13a). Silencing of 4 genes constituting the FAC core (for example, *PXN*, *SHC1*, *PITPNM2* and *PTK2B*), and 33 interactors, reduced WNV infection (Fig. 4b, c; Supplementary Fig. 4a, b)[25]. Reduction of WNV infection in *PITPNM2* silenced cells was also rescued with an siRNA resistant *PITPNM2* mutant (Supplementary Fig. 13b). Notably, only one core FAC component (PITPNM2) and 11 interactors reduced DENV infection (Fig. 4b, c). Among the nine HDAC components associated with WNV infection, four were required for DENV (Supplementary Fig. 13c). These examples indicate that WNV and DENV may have evolved different sensitivities in their interaction with host proteins, and this may be reflected in the differences in their biology.

In summary, this study portrays a comprehensive genome-scale map of human proteins and cellular pathways affecting the outcome of flavivirus–host cell interactions, and presents a potentially useful resource for further studies. Furthermore, these results may provide insights into the molecular differences in the pathogenesis of related flaviviruses, and reveal potential flaviviral therapeutic targets.

METHODS SUMMARY

RNAi screen. Any gene for which a minimum of two siRNAs reduced (HSF) or increased (HRF) the percentage of infected cells by ≥2-fold, and the fold change was ≥2 times the standard deviation (s.d.) of the percentage of control cells infected, was scored as a hit. Gene silencing that resulted in a cell number decrease ≥2 times the s.d. (of controls) was considered toxic and excluded.

Immunofluorescence assay sensitivity determination. As positive controls to determine whether the immunofluorescence assay (IFA) can detect changes in viral replication, the previously reported WNV replication impacting host gene *HMGCR* was silenced (Supplementary Fig. 1c, d); and to test whether IFA can detect changes in viral translation, host translation machinery was arrested by cycloheximide (Supplementary Fig. 1e). Anti-WNV siRNA was also used as a positive control (Supplementary Table 9). Results showed that the IFA was sensitive enough to detect changes in viral RNA translation, but not later stages such as replication.

Bioinformatic analysis. Genes were categorized using the PANTHER classification system[26]. Enrichment was analysed using the hypergeometric probability distribution. For tissue expression analysis, microarray data files were obtained from the Novartis GNF human expression atlas version 2 resource[27]. The protein network constructions used interaction data from the Human Protein Reference Database (HPRD)[28], the Biomolecular Interaction Network Database (BIND)[29] and the Ingenuity pathways database (Mountainview, CA), supplemented with functional information from the literature.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 23 March; accepted 26 June 2008.
Published online 6 August 2008.

1. Brinton, M. A. The molecular biology of West Nile Virus: A new invader of the western hemisphere. *Annu. Rev. Microbiol.* **56,** 371–402 (2002).
2. Chu, J. J. & Ng, M. L. Infectious entry of West Nile virus occurs through a clathrin-mediated endocytic pathway. *J. Virol.* **78,** 10543–10555 (2004).
3. Krishnan, M. N. *et al.* Rab 5 is required for the cellular entry of dengue and West Nile viruses. *J. Virol.* **81,** 4881–4885 (2007).
4. Brass, A. L. *et al.* Identification of host proteins required for HIV infection through a functional genomic screen. *Science* **319,** 921–926 (2008).
5. Ng, T. I. *et al.* Identification of host genes involved in hepatitis C virus replication by small interfering RNA technology. *Hepatology* **45,** 1413–1421 (2007).
6. Pelkmans, L. *et al.* Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis. *Nature* **436,** 78–86 (2005).
7. Davis, W. G., Blackwell, J. L., Shi, P. Y. & Brinton, M. A. Interaction between the cellular protein eEF1A and the 3′-terminal stem-loop of West Nile virus genomic RNA facilitates viral minus-strand RNA synthesis. *J. Virol.* **81,** 10172–10187 (2007).
8. Emara, M. M. & Brinton, M. A. Interaction of TIA-1/TIAR with West Nile and dengue virus products in infected cells interferes with stress granule formation and processing body assembly. *Proc. Natl Acad. Sci. USA* **104,** 9041–9046 (2007).
9. Hirsch, A. J. *et al.* The Src family kinase c-Yes is required for maturation of West Nile virus particles. *J. Virol.* **79,** 11943–11951 (2005).
10. Mackenzie, J. M., Khromykh, A. A. & Parton, R. G. Cholesterol manipulation by West Nile virus perturbs the cellular immune response. *Cell Host Microbe* **2,** 229–239 (2007).
11. Fredericksen, B. L., Smith, M., Katze, M. G., Shi, P. Y. & Gale, M. Jr. The host response to West Nile Virus infection limits viral spread through the activation of the interferon regulatory factor 3 pathway. *J. Virol.* **78,** 7737–7747 (2004).
12. Lim, J. *et al.* A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* **125,** 801–814 (2006).
13. Mashimo, T. *et al.* A nonsense mutation in the gene encoding 2′-5′-oligoadenylate synthetase/L1 isoform is associated with West Nile virus susceptibility in laboratory mice. *Proc. Natl Acad. Sci. USA* **99,** 11311–11316 (2002).
14. Samuel, M. A. *et al.* PKR and RNase L contribute to protection against lethal West Nile Virus infection by controlling early viral spread in the periphery and replication in neurons. *J. Virol.* **80,** 7009–7019 (2006).
15. Scherbik, S. V., Paranjape, J. M., Stockman, B. M., Silverman, R. H. & Brinton, M. A. RNase L plays a role in the antiviral response to West Nile virus. *J. Virol.* **80,** 2987–2999 (2006).
16. Tecle, T., White, M. R., Gantz, D., Crouch, E. C. & Hartshorn, K. L. Human neutrophil defensins increase neutrophil uptake of influenza A virus and bacteria and modify virus-induced respiratory burst responses. *J. Immunol.* **178,** 8046–8052 (2007).
17. Fujita, Y. *et al.* Hakai, a c-Cbl-like protein, ubiquitinates and induces endocytosis of the E-cadherin complex. *Nature Cell Biol.* **4,** 222–231 (2002).
18. Dell'Angelica, E. C. *et al.* AP-3: An adaptor-like protein complex with ubiquitous expression. *EMBO J.* **16,** 917–928 (1997).
19. Khor, R., McElroy, L. J. & Whittaker, G. R. The ubiquitin-vacuolar protein sorting system is selectively required during entry of influenza virus into host cells. *Traffic* **4,** 857–868 (2003).
20. Meusser, B., Hirsch, C., Jarosch, E. & Sommer, T. ERAD: The long road to destruction. *Nature Cell Biol.* **7,** 766–772 (2005).
21. Wakana, Y. *et al.* Bap31 is an itinerant protein that moves between the peripheral ER and a juxtanuclear compartment related to ER-associated degradation. *Mol. Biol. Cell* **19,** 1825–1836 (2008).
22. Schelhaas, M. *et al.* Simian Virus 40 depends on ER protein folding and quality control factors for entry into host cells. *Cell* **131,** 516–529 (2007).
23. Halestrap, A. P. & Price, N. T. The proton-linked monocarboxylate transporter (MCT) family: Structure, function and regulation. *Biochem. J.* **343,** 281–299 (1999).
24. Chiu, M. W., Shih, H. M., Yang, T. H. & Yang, Y. L. The type 2 dengue virus envelope protein interacts with small ubiquitin-like modifier-1 (SUMO-1) conjugating enzyme 9 (Ubc9). *J. Biomed. Sci.* **14,** 429–444 (2007).
25. Hanks, S. K., Ryzhova, L., Shin, N. Y. & Brabek, J. Focal adhesion kinase signaling activities and their implications in the control of cell survival and motility. *Front. Biosci.* **8,** d982–d996 (2003).
26. Mi, H. *et al.* The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.* **33,** D284–D288 (2005).
27. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA* **101,** 6062–6067 (2004).
28. Mishra, G. R. *et al.* Human protein reference database — 2006 update. *Nucleic Acids Res.* **34,** D411–D414 (2006).
29. Bader, G. D., Betel, D. & Hogue, C. W. BIND: The Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31,** 248–250 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The human genome RNAi library was made available through the support of the New England Regional Center of Excellence in Biodefense and Emerging Infectious Disease (U54AI057159). The screening was performed at the ICCB-Longwood screening facility (Harvard Medical School). We thank B. Lindenbach for suggestions and Y. Benita for illustrations. We thank L2 Diagnostics for providing the anti-WNV antibody. This work was supported by the NIH. A.N. is supported by a fellowship award from the Crohn's and Colitis Foundation of America. R.J.X. is supported by the NIH (AI062773) and by CCIB development funds. F.D.G was supported by an NIH training grant in Emerging and Tropical Infectious Diseases (AI07526); portions of this work were supported by a grant from NIAID to P.W.M. through the WRCE (NIH U54 AI057156). E.F. and S.J.E. are Investigators of the Howard Hughes Medical Institute.

Author Contributions M.N.K., H.A. and E.F. designed the experiments; M.N.K., B.S., E.F. and R.A. performed the screen; M.N.K., B.S., R.A.K., A.L.B., S.J.E. and H.A. analysed the data; M.N.K. and H.S. performed validations; P.D.U. designed microscopy; F.D.G. and P.W.M. designed replicon experiments; S.L. provided PITPNM2 cDNA; A.N. and R.J.X. performed bioinformatics analyses; and M.N.K., H.A., A.N., R.J.X. and E.F. co-wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to E.F. (erol.fikrig@yale.edu).

METHODS

WNV RNAi screen and candidate protein selection criteria. A library of 21,121 siRNA pools targeting human genome (Dharmacon siARRAY siRNA Library, Human Genome, G-005000-05, Thermo Fisher Scientific) was used. For both the primary and validation screens, HeLa cells (384-well format) were transfected (using Dharmafect 1, Dharmacon) in duplicates with siRNA (50 nM) for 72 h, infected for 24 h with West Nile virus (WNV strain 2471) or 30 h with dengue virus 2 (DENV New Guinea C strain), fixed in 4% paraformaldehyde, immunostained with antibodies detecting viral E-proteins (TRITC labelled, anti-WNV-E antibody developed in horse, or monoclonal anti-DENV-E, Chemicon), and imaged by fluorescence microscopy (Molecular Devices, 4× magnification) using a TRITC filter for virus and DAPI filter for nuclei. Infection was done at an m.o.i. of 0.3 for both WNV and DENV. Generally, infection was in the range of 20–30% for both WNV and DENV. As positive control of infection reduction due to gene silencing, endosomal proton pump vATPase was silenced. Cell number per well was in the range 7,000–9,000. The percentage infection was relatively linear in the cell number range in which the screen was performed (Supplementary Fig. 1f). Each 384-well plate had additional control wells with a non-targeting control siRNA (siCONTROL non-targeting siRNA, Dharmacon) (for determining the general effect of siRNA transfection on infection), siRNA targeting *PLK-1* whose silencing kills the cells (for determining general knockdown efficiency), fluorescently labelled non-targeting control siRNA (for determining transfection efficiency), and wells with neither transfection reagent nor siRNA. Quantification of the effect of gene silencing on viral infection was done using the software Metamorph (Molecular Devices), which counted cells that were immuno-stained versus non-stained for virus antigen. Based on the infection kinetics and infection inhibition by the silencing of a host gene known to be required for the infection of both WNV and DENV (vATPase, Supplementary Fig. 1b), we defined an infection reduction of twofold or greater at 24 h for WNV or 30 h for DENV as the threshold for hit selection. Silencing of vATPase resulted in a reduction of infection of 2.9 ± 0.3 fold compared to the controls for WNV or 2.7 ± 0.4 for DENV (Supplementary Fig. 1b).

Cell lines and virus propagation. Gene silencing and infection studies were done on low passage HeLa (ATCC no. CCL-2.1) cells maintained in DMEM supplemented with 10% fetal bovine serum. West Nile virus (strain 2471, gift of J. Anderson), and dengue 2 virus (New Guinea C strain, gift of A. de Silva) viruses were grown on vero (ATCC no. CRL-1586) or C6/36 (ATCC no. CRL-1660) cells, respectively.

Gene knockdown verification, RNAi resistant mutant generation and phenotype rescue. For the quantitation of the target transcript reduction, pooled siRNAs corresponding to the tested genes were transfected (50 nM) to cells (or cells pre-transfected with cDNAs of genes) in 48-well plates for 3 days, total RNA was isolated using the RNeasy kit (Qiagen), and cDNA was prepared using the iScript kit (Biorad). Quantitative PCR (qPCR) was performed by using Sybergreen reagent (Biorad). The primers used were given in the Supplementary Table 7. To generate RNAi resistant variants of genes, four silent mutations each was introduced into those sequences of *DERL2* and *PITPNM2* where siRNA binds (in the expression vector pCDNA6.2 with V5 tag), using QuickChange Mutagenesis kit (Strategene) (Supplementary Table 7 shows the mutagenesis primer sequences). HeLa cells transfected separately with wild type or mutant copies of the genes were selected for 8 days using blasticidin, treated with siRNA for 72 h, and either WNV infection assay or western blot (for knock down and rescue verification) was performed. Six random fields of fluorescent images (10× objective, Zeiss Axiovert 200M) of WNV infected mutant versus wild type gene expressing cells were counted to quantify and assess the rescue of viral infection by mutant genes. For western blot, cells were lysed in 1% 50 mM Tris-HCL, 150 mM NaCl and Triton X-100. Western blot was performed to verify the extent of knockdown and rescue of *DERL2* and *PITPNM2* using anti-V5 antibodies (Invitrogen). Anti-CBLL1 antibody was obtained from Abcam.

Cytotoxicity. Cytotoxic effects of gene silencing and MG132 treatments were determined using LDH release assay kit (Roche). Supernatants of gene silenced cells were harvested at 3 days post-transfection or 1–24 h post treatment for compounds, and assayed for LDH release according to manufacturer's protocol.

Viral RNA transfection and secretion studies. Two kinds of studies were done using viral genomic RNA transfection: (1) determination of the effect of gene silencing on viral RNA translation, and (2) determination of the effect of gene silencing on progeny virion secretion. The RNA of a subgenomic replicon of WNV (lacking complete genes for the capsid, pre-membrane, and envelope proteins) was used for viral translation studies[30], while a full length viral genome was used for viral secretion studies[31]. The viral RNA was prepared as described previously[31]. The viral RNA was transiently transfected into HeLa cells by electroporation, after gene knock down with siRNA for 3 days. Mouse hyperimmune ascitic fluid against WNV was used for the immunofluorescence of replicon transfected cells, after 14 h of transfection. For the viral secretion assay, the culture supernatants were collected at 24 h from HeLa cells electroporated with (400 ng) full length WNV genomic RNA for ERAD silenced cells. As positive control for inhibition of WNV secretion, brefeldin A ($10\ \mu g\ ml^{-1}$) was used, by adding 12 h post infection. To study the viral release from *MCT4* silenced cells, supernatants were collected at 12 h and 24 h (post-infection), and performed a plaque formation assay. For anti-WNV siRNA (siRNA sequence is given in Supplementary Table 9) treatment of replicon silenced cells, cells are first transfected with anti-WNV siRNA for 6 h, followed by electropration of replicon, and fixed after 30 h for IFA.

Inhibitor studies. HeLa cells were treated with 15 μM MG132 (Biomol) or $700\ \mu g\ ml^{-1}$ cycloheximide (Sigma) (dissolved in DMSO) or DMSO for various time periods as described in the text or figure. For determining the role ubiquitination in viral internalization, HeLa cells were pre-treated with MG132 for 1 h, TRITC-WNV was added (m.o.i. ≈ 100), incubated for 1–4 h at 37 °C, fixed with 4% paraformaldehyde, and confocal microscopy was performed. For determining the post-entry requirements of ubiquitination, the virus was inoculated (m.o.i. ≈ 0.3), incubated at 37 °C, and MG132 was added at different time points. The cells were fixed and immunostained after 18 h. The final DMSO concentration was no more than 0.2% of the total culture medium.

Vesicular stomatitis virus (VSV) and HIV studies. The human immunodeficiency virus (VSV-G carrying pHXBGFP-IRES-nef, an infectious molecular clone expressing GFP; gift of P. Shankar) infection experiments were done by infecting gene silenced cells at an m.o.i. of 0.3 for 24 h, followed by fluorescent imaging (10×, Zeiss) to quantify the percentage infection. For the VSV experiments, cells were pre-treated for 1 h with either DMSO or 15 μM MG132, followed by infection with VSV expressing GFP (m.o.i. ≈ 0.5) for 12 h. GFP-positive VSV cells were quantified by flow cytometry.

WNV entry and colocalization studies. Modification of a previous protocol was used for these studies[32]. Purified virus was exchanged into phosphate buffered saline (PBS, pH 7.4) through repeated cycles of concentration by centrifugation (800*g*) and dilution with PBS, using 15 ml ultrafiltration tubes (10kD, Amicon). The virus in PBS (equivalent to 0.5 mg per ml protein) was incubated with tetramethyl rhodaminyl isothiocyanate (TRITC, Pierce Biotechnology) ($0.3\ mg\ ml^{-1}$, in dimethyl formamide) for 1 h at room temperature. After removal of excess dye, labelled virus (WNV–TRITC) was immediately used for experiments. Labelling did not abolish viral infectivity. Infectious entry of WNV–TRITC was sensitive to the vATPase inhibitor bafilomycin, and colocalized with Rab5 labelled compartments, similar to the entry mechanism of unlabelled WNV (Supplementary Fig. 14a, b). For colocalization imaging, Rab5–GFP or Rab7–GFP transduced HeLa cells were used (gift from T. Dragic). For the entry assay (or Rab5/7–GFP colocalization experiments), cells in 48-well culture plates were transfected with siRNA for 48 h, and re-plated onto glass slide bottom chambers (MatTek) in DMEM (with 5% serum and 20 mM Hepes, pH 7.4). After further 24 h, WNV–TRITC (m.o.i. ≈ 100, to capture sufficient events) was added to the cells and allowed to bind for 1 h at 4 °C, and cells were shifted to 37 °C for different time periods, fixed with 4% paraformaldehyde, and confocal imaging was performed, on a LSM 510 confocal microscope equipped with a Zeiss axiovert 100 M base, using 100× oil objective (Zeiss MicroImaging). Z-stack imaging was done at 0.5 μm sections. Virus particles within the cells were counted using the software velocity and ImageJ.

Enrichment analysis of biological process and molecular function categories. Genes were classified into biological process and molecular function terms using the PANTHER system. To assess the statistical enrichment or over-representation of these categories for the set of hits relative to the global set of genes examined in the RNAi screen, *P* values were computed using the hypergeometric probability distribution, which was implemented in the R language (http://www.r-project.org/). The hypergeometric distribution describes the probability of finding *s* genes associated with a particular category, in a set of *g* genes essential for WNV infection (identified from the RNAi screen), given that there are *S* genes associated with that same category in the global set of *G* genes examined in the genome-wide RNAi screen. For each category *c*, and the list of genes *l*, the *P* value was calculated as:

$$P(c, l) = 1 - \sum_{k \in \{0, 1, \ldots, s\}} [C(g, k)\ C(G - g, S - k) / C(G, S)]$$

The binomial coefficient is of the form $C(n, r)$. A *P* value <0.05 was considered significant. Categories assigned with at least 10 genes are displayed in Fig. 1b, c. A similar approach was used to examine over-representation in Fig. 4a, except that the assessment of enrichment for biological process categories was made relative to their representation among all HSFs identified.

Analysis of gene expression across 79 tissues. Microarray data files were obtained from the Novartis GNF human expression atlas version 2 resource, and expression values of 33,689 probe sets from the HG-U133A (Affymetrix) platform and the GNF1H custom chip were analysed. The data set was normalized using global median scaling, and we filtered the data by excluding from the analysis those probe sets with 100% 'absent' calls (MAS5.0 algorithm) across all 79 tissues. The data set was further filtered by setting a minimum threshold value >20 in at least one sample for each probe set and a maximum-mean expression value >100. Hierarchical clustering (centroid linkage method) was performed with Cluster 3.0 using Pearson's correlation as the similarity metric[33]. *Z*-score transformation was applied to each probeset across all arrays before generating 'heatmaps' for visualization using TreeView[34].

Constructing human protein interaction network. The protein network construction used protein interaction data obtained from the Human Protein Reference Database (HPRD), Biomolecular Interaction Network Database (BIND), Ingenuity pathways database (Mountainview, CA) and functional information from the literature. The network uses graph theoretical representations in which components (gene products) are depicted as nodes and interactions between components as edges. Graph layout descriptions were written in the Dot language, which implements a multi-dimensional scaling heuristic and uses an iterative solver (Newton-Raphson algorithm) that searches for low-energy configurations to optimize the graph layout when creating a virtual physical model (Spring model) for visualization.

30. Scholle, F. & Mason, P. W. West Nile virus replication interferes with both poly(I:C)-induced interferon gene transcription and response to interferon treatment. *Virology* **342,** 77–87 (2005).
31. Rossi, S. L., Zhao, Q., O'Donnell, V. K. & Mason, P. W. Adaptation of West Nile virus replicons to cells in culture and use of replicon-bearing cells to probe antiviral action. *Virology* **331,** 457–470 (2005).
32. Helenius, A., Kartenbeck, J., Simons, K. & Fries, E. On the entry of Semliki forest virus into BHK-21 cells. *J. Cell Biol.* **84,** 404–420 (1980).
33. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95,** 14863–14868 (1998).
34. Saldanha, A. J. Java Treeview — extensible visualization of microarray data. *Bioinformatics* **20,** 3246–3248 (2004).

LETTERS

T-cell-expressed proprotein convertase furin is essential for maintenance of peripheral immune tolerance

Marko Pesu[1], Wendy T. Watford[1], Lai Wei[1], Lili Xu[2], Ivan Fuss[2], Warren Strober[2], John Andersson[3], Ethan M. Shevach[3], Martha Quezado[5], Nicolas Bouladoux[4], Anton Roebroek[6], Yasmine Belkaid[4], John Creemers[7] & John J. O'Shea[1]

Furin is one of seven proprotein convertase family members that promote proteolytic maturation of proproteins[1]. It is induced in activated T cells and is reported to process a variety of substrates including the anti-inflammatory cytokine transforming growth factor (TGF)-β1 (refs 2–4), but the non-redundant functions of furin versus other proprotein convertases in T cells are unclear. Here we show that conditional deletion of furin in T cells allowed for normal T-cell development but impaired the function of regulatory and effector T cells, which produced less TGF-β1. Furin-deficient T regulatory (T_{reg}) cells were less protective in a T-cell transfer colitis model and failed to induce Foxp3 in normal T cells. Additionally, furin-deficient effector cells were inherently overactive and were resistant to suppressive activity of wild-type T_{reg} cells. Thus, our results indicate that furin is indispensable in maintaining peripheral tolerance, which is due, at least in part, to its non-redundant, essential function in regulating TGF-β1 production. Targeting furin has emerged as a strategy in malignant and infectious disease[5,6]. Our results suggest that inhibiting furin might activate immune responses, but may result in a breakdown in peripheral tolerance.

Proprotein convertases are important for the maturation of multiple cellular substrates, but the identification of non-redundant, bona fide enzyme/substrate pairs is lacking. Furthermore, germline deletion of furin is embryonically lethal, limiting our understanding of the physiological function of this proprotein convertase[7]. To investigate furin's role in T cells, we crossed mice homozygous for floxed *fur* alleles[8] with mice expressing Cre recombinase under the control of the *CD4* promoter (designated CD4cre-fur$^{f/f}$ mice). T-cell-specific deletion of *fur* resulted in a viable mouse in which furin messenger RNA and protein were virtually absent in both CD4 and CD8 compartments (Supplementary Fig. 1). Furin-deficient T cells underwent normal thymic development as evidenced by normal absolute T-cell numbers (data not shown), ratios of thymic subsets and T-cell antigen receptor (TCR)-β rearrangement (Fig. 1a). In young animals, the absolute numbers of T cells, proportions of CD4$^+$ and CD8$^+$ T cells and TCR Vβ subsets in peripheral lymphoid organs (spleen and lymph nodes) were also not significantly different from the fur$^{f/f}$ littermate controls (data not shown and Supplementary Fig. 2). In addition, partial deletion of the Vβ5$^+$ subset in peripheral CD4$^+$ but not CD8$^+$ T cells was evident in the absence of furin, suggesting that negative selection of thymocytes is intact (Supplementary Fig. 2)[9]. Thus, deletion of furin at the double-positive stage of T-cell development did not appear to have major developmental consequences. Intriguingly though, the numbers of thymic, natural T regulatory (T_{reg}, CD4$^+$Foxp3$^+$) cells were found to be significantly elevated in CD4cre-fur$^{f/f}$ animals (Fig. 1b).

Although deleting furin in double-positive thymocytes did not grossly affect T-cell development, furin deficiency in T cells was associated with increased numbers of activated, memory-like CD4$^+$CD44^{hi}CD62L$^-$ and CD8$^+$CD44^{hi}CD122$^+$ T cells in the periphery even in 7- to 9-week-old mice (Fig. 1c). To gain more insight into the biological consequence of the absence of furin in T cells, we performed microarray analysis on sorted naive, CD4$^+$CD44^{low}CD62L$^+$ CD4cre-fur$^{f/f}$ and littermate fur$^{f/f}$ T cells (greater than 98% purity). Although the cells were isolated based on their naive phenotype, the absence of furin was associated with the upregulation of several genes typically associated with T-cell activation, including *Fos*, *Jun* and *Ifng* (Supplementary Fig. 3). Moreover, upon activation, furin-deficient T cells were observed to produce greater amounts of Th1- (interferon (IFN)-γ) and Th2 (interleukin (IL)-4 and IL-13)-type cytokines, less IL-2 and equivalent levels of TNF or IL-17 (Fig. 1d and Supplementary Fig. 4).

At approximately 6 months of age, CD4cre-fur$^{f/f}$, but not littermate fur$^{f/f}$ or CD4cre-fur$^{+/+}$, mice became overtly ill, at which point they developed a progressive wasting disease characterized by weight loss, ruffled hair and hunched appearance. Gross pathological examination of the large intestine and stomach of CD4cre-fur$^{f/f}$ mice revealed macroscopic evidence of inflammation and fibrosis (Fig. 2a). Mesenteric lymph nodes were enlarged, but obvious splenomegaly was rarely observed. Histologically, mice had severe inflammatory bowel disease characterized by dense chronic inflammation with reactive epithelial atypia and architectural distortion; scattered neutrophils were also observed. Nodules of lymphoid infiltrates were also noted in the liver, lung and kidney, and CD4cre-fur$^{f/f}$ mice were also found to have high levels of anti-nuclear and anti-DNA antibodies, indicative of systemic autoimmune disease (Fig. 2b, c, and Supplementary Figs 4 and 5). Analysis of serum cytokines in CD4cre-fur$^{f/f}$ mice revealed elevated levels of circulating pro-inflammatory cytokine IL-6, the hallmark Th1 and Th2 cytokines IFN-γ and IL-13, respectively, and lower levels of the anti-inflammatory cytokine IL-10 (Fig. 2d). In addition, there was secondary activation of B cells, as evidenced by elevated levels of the serum immunoglobulins

[1]Molecular Immunology and Inflammation Branch, National Institute for Arthritis, Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA. [2]Mucosal Immunity Section, Laboratory of Host Defenses, [3]Cellular Immunology Section, Laboratory of Immunology, [4]Mucosal Immunology Unit, Laboratory of Parasitic Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA. [5]Laboratory of Pathology, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [6]Experimental Mouse Genetics, [7]Laboratory for Biochemical Neuroendocrinology, K.U. Leuven and V.I.B., B-3000 Leuven, Belgium.

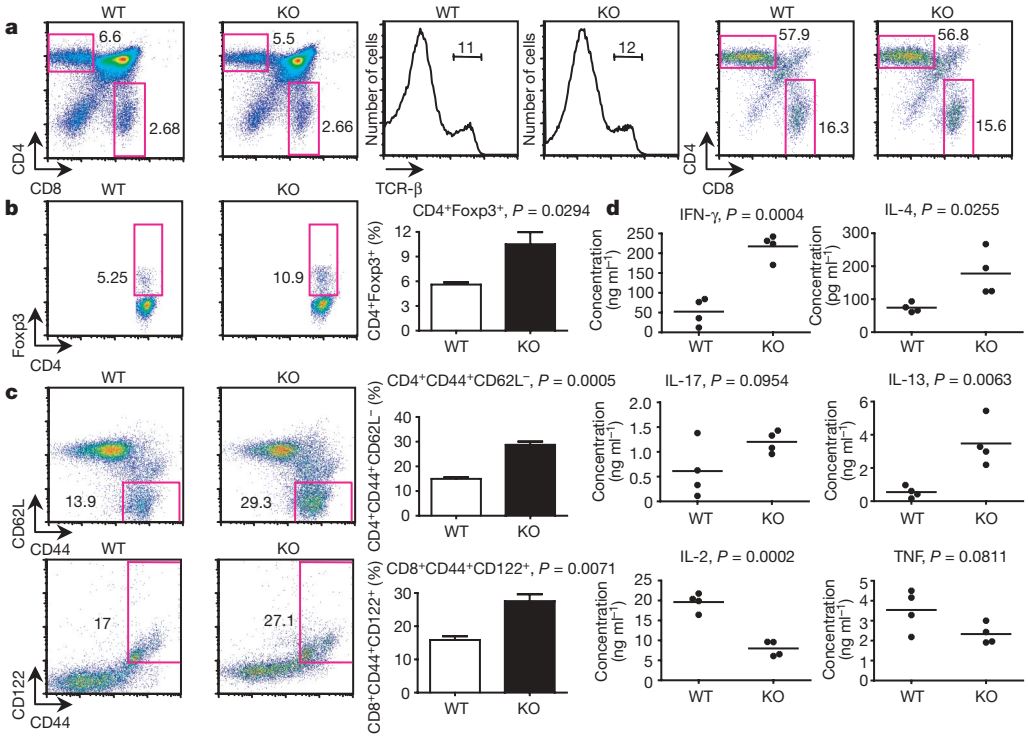


Figure 1 | Normal thymic T-cell development, but activated/memory phenotype of peripheral T cells in CD4cre-fur$^{f/f}$ mice. **a**, **b**, Proportions of CD4$^+$, CD8$^+$, TCR-β$^+$ and Foxp3$^+$ thymocytes in 8-week-old mice. Panels on the right show CD4$^+$/CD8$^+$ proportions in TCR-β rearranged cells. Three mice per group were analysed. Representative flow cytometry blots and plotted mean values are shown. **c**, Activated/memory splenic T cells in 7- to 9-week-old mice. Representative flow cytometry blots and plotted mean values are shown ($n = 3$ per group). **d**, Cytokine production. Mesenteric lymph node cells were stimulated with plate-bound CD3 and soluble CD28 antibodies for 48 h ($n = 4$, 9-week-old CD4cre-fur$^{f/f}$ and fur$^{f/f}$ mice). Error bars, s.e.m. KO, knockout; WT, wild type.

IgG1, IgG2 and IgE (Fig. 2e). In the gut and mesenteric lymph nodes of older, autoimmune CD4cre-fur$^{f/f}$ animals, we also observed expansion of CD4$^+$ and CD8$^+$ effector cells, as well as upregulation of the activation marker CD69 on CD4$^+$ effector cells (Fig. 3a, b). Together with the evidence of T-cell activation, autoantibody production and hyperproduction of Th1 and Th2 cytokines, these data suggest that CD4cre-fur$^{f/f}$ mice had a breakdown in immune tolerance.

Regulatory T cells are critical components of T-cell-dependent peripheral tolerance[10]. We found that mice in which furin was deleted in T cells had elevated numbers and proportions of peripheral CD4$^+$Foxp3$^+$ cells in the small intestine and mesenteric lymph nodes (Fig. 3a, d). Production of the anti-inflammatory cytokine TGF-β1 by T_{reg} and effector T cells is also critical in the maintenance of peripheral tolerance and prevention of autoimmune disease[11–13]. TGF-β1 is initially synthesized as a proprotein that requires multiple steps to generate the mature, biologically active cytokine[14]. The proprotein is enzymatically cleaved to generate an amino-terminal latency associated peptide, but this product remains non-covalently associated with TGF-β1, keeping it in a biologically inactive state. Previous studies with recombinant furin have argued for a role as a pro-TGF-β1 converting enzyme, although other proprotein convertase family members have also been reported to have this activity[2]. We first explored furin's criticality in the production of biologically active TGF-β1 by assessing whether furin-deficient T cells generated normal levels of TGF-β1 *in vitro*. To this end, we measured the secretion of TGF-β1 by CD4$^+$CD25$^-$ and CD4$^+$CD25$^+$ T cells and found that furin-deficient T cells secreted considerably reduced levels of activated TGF-β1 (Fig. 3c). We confirmed this deficit by measuring TGF-β1 by enzyme-linked immunosorbent assay (ELISA) and western blot using antibodies that selectively detected only active TGF-β1 and not the unprocessed cytokine (Supplementary Fig. 6c, d). In contrast, the levels of surface-expressed latency associated peptide and total *Tgfb1* mRNA were not reduced in furin-deficient T cells, consistent with normal production of pro-TGF-β1 (Supplementary Fig. 6a, b and data not shown).

Foxp3 can be induced in naive CD4$^+$ T cells by addition of exogenous TGF-β1 (ref. 15). Moreover, it has been recently found that TGF-β1 produced by T_{reg} cells also upregulates Foxp3 expression in co-cultures with CD4$^+$Foxp3$^-$ cells[16]. As shown in Supplementary Fig. 7, addition of TGF-β1 or wild-type T_{reg} cells induced the expression of Foxp3 in both wild-type and furin-deficient CD4$^+$ naive T cells. However, furin-deficient T_{reg} cells were defective in their ability to upregulate Foxp3 expression in normal CD4$^+$ T cells *in vitro* (Supplementary Fig. 7). This last finding is consistent with the reduced capacity of furin-deficient T cells to produce normal levels of biologically active TGF-β1; however, it is also clear that furin-deficient T cells can respond to exogenously added TGF-β1.

To assess whether bioactive TGF-β1 was reduced in the absence of furin *in vivo*, we first examined the expression of the integrin CD103 on T lymphocytes. This integrin has been linked to T-cell gut homing and retention within epithelial compartments and is known to be induced by TGF-β1 (refs 11, 17). We found that numbers of CD4$^+$CD103$^+$Foxp3$^-$ cells in the lamina propria and intraepithelial compartments were reduced in the small intestines of CD4cre-fur$^{f/f}$ mice compared with wild-type mice (Fig. 3d).

T-cell-derived TGF-β1 is essential for suppressing autoimmunity, and TGF-β1-deficient T_{reg} cells cannot suppress the inflammatory bowel disease caused by homeostatic expansion of effector T cells[12]. To assess their function, furin-deficient naive CD4$^+$CD25$^-$CD45Rb^{hi} cells (effectors) were injected into T-cell-deficient hosts alone, or in combination with CD4$^+$CD25$^+$ T_{reg} cells. Mice that received either wild-type or furin-deficient effector cells with no T_{reg} cells lost weight and developed severe gut inflammation macroscopically and histologically (Fig. 4A–C). Furin-deficient effector T cells exhibited a more aggressive phenotype, as evidenced by a significant increase in the absolute number of knockout CD4$^+$ cells in mesenteric lymph nodes (Fig. 4D).

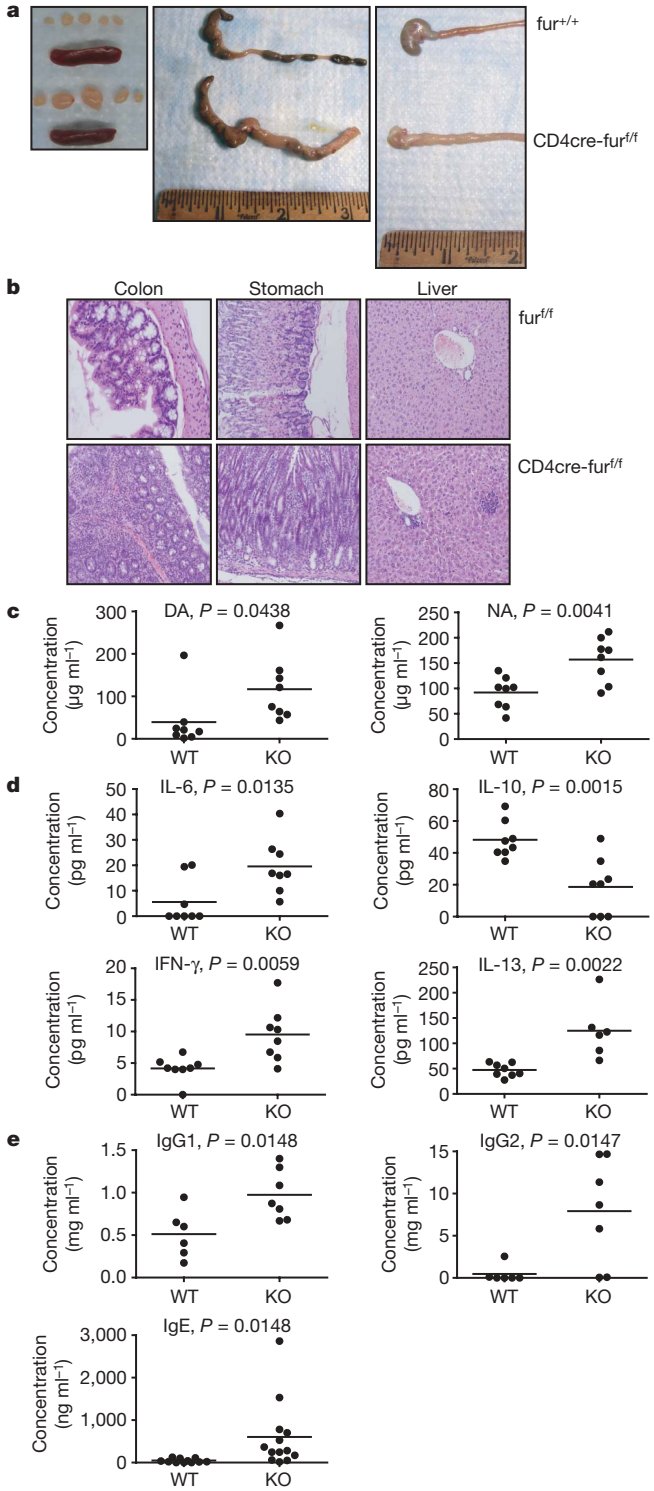


Figure 2 | Development of age-related autoimmunity in CD4cre-fur$^{f/f}$ mice. **a**, Lymphoid organs, colon and stomach/duodenum of CD4cre-fur$^{f/f}$ and age-matched wild-type animals (6 months). **b**, Haematoxylin- and eosin-stained sections of colon, stomach and liver (6-months-old, CD4cre-fur$^{f/f}$ and fur$^{f/f}$ mice). **c**, Anti-double-stranded DNA (DA) and nuclear antibody (NA) titres in CD4cre-fur$^{f/f}$ animals compared with age-matched wild-type and fur$^{f/f}$ animals ($n = 8$, 5–7 months). **d**, Serum cytokines in CD4cre-fur$^{f/f}$ animals compared with age-matched wild-type and fur$^{f/f}$ animals ($n = 6$–8, 5–7 months). **e**, ELISA for serum immunoglobulins of CD4cre-fur$^{f/f}$ animals compared with age-matched wild-type and fur$^{f/f}$ animals ($n = 6$–13, 5–7 months). KO, knockout; WT, wild type.

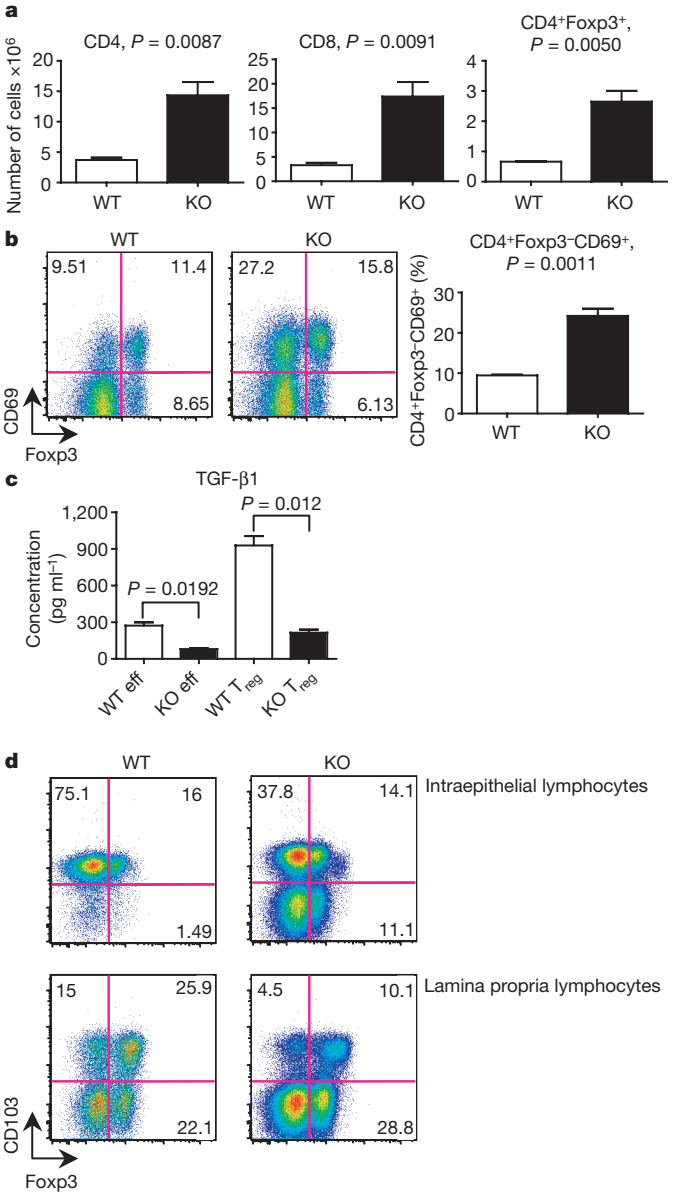


Figure 3 | Deletion of furin in T cells results in T-cell expansion/activation, and impairs TGF-β1 production and CD103 expression. **a**, **b**, Absolute CD4$^+$Foxp3$^-$, CD4$^+$Foxp3$^+$ and CD8$^+$ cell numbers and the proportion of CD4$^+$Foxp3$^-$CD69$^+$ cells in the mesenteric lymph nodes of CD4cre-fur$^{f/f}$ animals were compared with age-matched wild-type and fur$^{f/f}$ animals ($n = 3$, 5–6 months). **c**, TGF-β1 production. Purified CD4cre-fur$^{f/f}$, fur$^{f/f}$ CD4$^+$CD25$^-$ and CD4$^+$CD25$^+$ cells were activated with plate-bound CD3 and soluble CD28 antibodies; TGF-β1 was measured by ELISA in duplicate and a representative of three experiments is shown. **d**, Expression of Foxp3 and CD103 in lamina propria and intraepithelial CD4cre-fur$^{+/+}$ and CD4cre-fur$^{f/f}$ CD4$^+$ T cells; shown is a representative flow cytometry plot of two mice per group analysed (8–10 months). Error bars, s.e.m.; eff, effector; KO, knockout; WT, wild type.

Consistent with previous reports, transfer of wild-type T_{reg} cells with wild-type effectors prevented the wasting disease, gut inflammation, CD4$^+$ cell expansion and IFN-γ production (Fig. 4 and Supplementary Fig. 8)[18]. However, furin-deficient T_{reg} cells were found to be significantly impaired in their ability to prevent gut pathology and weight loss, as well as inhibiting homeostatic proliferation and cytokine production by effector CD4$^+$ cells. Intriguingly, furin-deficient effectors could not be completely suppressed by wild-type T_{reg} cells (Fig. 4C and Supplementary Fig. 8). During homeostatic expansion of naive CD4$^+$ cells, spontaneous conversion of

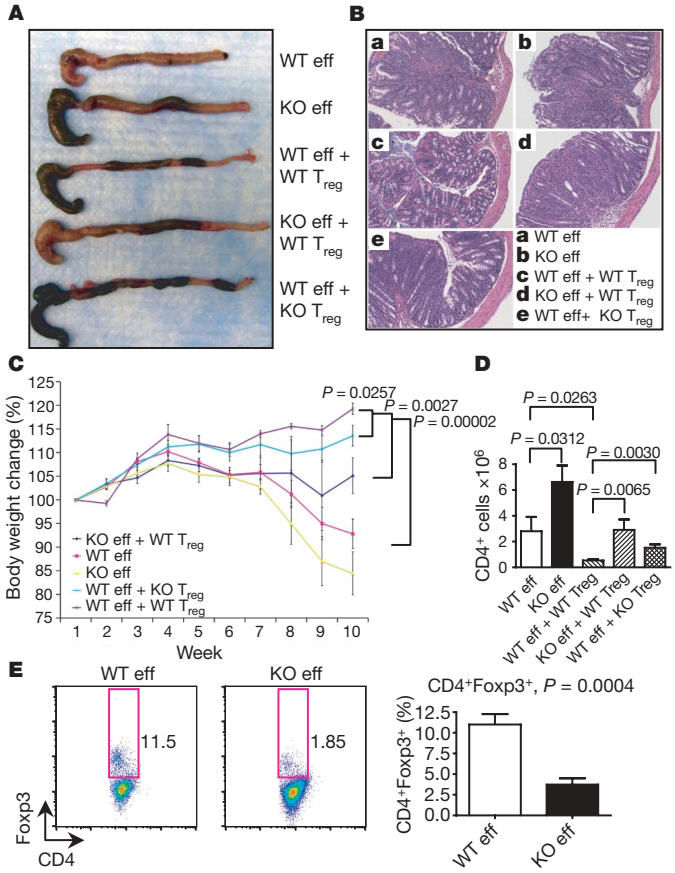


Figure 4 | Furin deficiency results in increased T effector cell aggressiveness and impaired T regulatory cell protection in suppressing colitis. Purified wild-type (WT) or CD4cre-fur$^{f/f}$ (KO) $CD4^+CD25^-CD45Rb^{hi}$ naive T cells were transferred alone or in combination with wild-type or furin-deficient $CD4^+CD25^+$ T regulatory cells into TCR-$\alpha^{-/-}$ recipients. Mice were analysed on week 10 ($n = 4$ or 5 per group). **A**, Representative images of colons. **B**, Representative colon histology. **C**, Change in body weight during the experiment. **D**, Absolute $CD4^+$ cell numbers in mesenteric lymph nodes. **E**, Spontaneous conversion of adoptively transferred naive $CD4^+CD25^-CD45Rb^{hi}$ cells into $CD4^+Foxp3^+$ cells. Representative flow cytometry blots and plotted mean values are shown. In **E**, data are pooled from two identical experiments. Error bars, s.e.m.; eff, effector; KO, knockout; WT, wild type.

$CD4^+$ T cells to $Foxp3^+$ cells occurs in a TGF-β1-dependent manner (typically 10%–15%)[19,20]. We found that adoptive transfer of naive furin-deficient $CD4^+$ cells into T-cell-deficient hosts resulted in the generation of fewer $Foxp3^+$ cells than wild-type cells, perhaps owing to lack of autocrine TGF-β1 production (Fig. 4E). Lack of spontaneous conversion may contribute to the enhanced aggressiveness of furin-deficient effectors *in vivo*.

Previous *in vitro* studies using recombinant proteins have suggested that furin processes a variety of substrates[4]. Fundamental, non-redundant roles in cell biology are suggested by the early lethality of *fur* germline knockout. However, it was notable that T-cell-specific furin-deficient animals have grossly normal T-cell development, arguing that furin is not required for a plethora of substrates, consistent with the observed partial redundancy of furin in the liver[8].

Rather, selectively deleting furin in T cells, like deleting TGF-β1 in T cells[12], resulted in autoimmunity and inflammatory bowel disease characterized by elevated numbers of thymic/natural T regulatory cells, hyperproduction of both Th1 and Th2 cytokines and expansion of $CD4^+$ and $CD8^+$ cells. Similar to TGF-β1-deficient T cells, *in vitro* suppressive activity[21] was not impaired (Supplementary Fig. 9), but *in vivo* suppressive activity was reduced, and effector T cells were more aggressive. Thus, CD4cre-fur$^{f/f}$ mice largely phenocopy the abnormalities seen in CD4cre-tgfb$^{f/n}$ mice. The phenotype of CD4cre-fur$^{f/f}$ mice also mimics the pathology seen in mice in which another molecule involved in the activation of TGF-β1, integrin $\alpha_v\beta_8$, is deleted in dendritic cells[22]. It is noteworthy, however, that unlike TGF-β1-deficient T_{reg} cells, CD4cre-fur$^{f/f}$ $CD4^+CD25^+$ cells are partly functional *in vivo* and furin-deficient effector cells are more resistant to T_{reg} suppression (Fig. 4). Undoubtedly, furin has other substrates in T cells, but our results indicate that furin is a critical proprotein convertase *in vivo* for the proper endoproteolytic processing of TGF-β1. Although other proprotein convertases have also been reported to cleave TGF-β1 *in vitro*, they evidently cannot replace furin in T cells. Although our data argue for a role of T-cell-expressed furin in the maintenance of peripheral tolerance without obvious effects on thymic selection, more work on furin's role in T-cell development and central tolerance is clearly warranted.

Our results have additional implications. Furin activity has been linked to the pathogenesis of several diseases including metastatic cancers, cystic fibrosis and infectious diseases[4,5,23,24]. Consequently, furin inhibitors have been proposed as possible therapies for such diseases. However, our findings suggest that interfering with furin activity might have the unexpected consequence of promoting autoimmunity. In principle, this might be beneficial in that it might boost T-cell-mediated immune responses and be advantageous in treating cancer and infections.

METHODS SUMMARY

Mice. Mice that expressed floxed *fur*[8] alleles were backcrossed six times with C57BL/6 mice. Wild-type, CD4-Cre and TCR-$\alpha^{-/-}$ mice on C57BL/6 background were from Taconic. Mice were maintained and housed under pathogen-free conditions in accordance with the National Institutes of Health Animal Care and Use Committee.

Cell purification and flow cytometry. Cells were purified by magnetic separation (Miltenyi) and sorted with a MoFlo cell sorter (Dako). Flow cytometry was performed with FACSCanto or FACSCalibur instruments (Becton Dickinson) and data were analysed using FlowJo software (Treestar).

Cytokine and antibody measurements. Mesenteric lymph node T cells (1×10^6 ml^{-1}) were activated with plate-bound anti-CD3 (10 μg ml^{-1}) and soluble anti-CD28 (2 μg ml^{-1}; BD Pharmingen). Cytokine levels in supernatants and sera were determined with IL-13 and IL-17 ELISA (R&D Systems) or with Cytometric Bead Array (BD Pharmingen). Serum auto-antibodies and immunoglobulins were determined by ELISA (Alpha Diagnostic International). TGF-β1 production by purified $CD4^+CD25^-$ and $CD4^+CD25^+$ cells was measured after two rounds of activation using a multispecies TGF-β1 ELISA kit (Invitrogen).

Gut cell preparation. Intraepithelial lymphocytes were purified from the small intestine using mechanical separation and 30% Percoll centrifugation; lamina propria cells were purified as described[25].

***In vivo* suppression assay.** The T-cell suppression assay was performed as previously described with small modifications[18]. Briefly, TCR-$\alpha^{-/-}$ mice were injected intravenously with purified wild-type or furin-deficient naive $CD4^+CD25^-CD45Rb^{hi}$ cells with or without wild-type or furin-deficient $CD4^+CD25^+$ T_{reg} cells. Mice were monitored for signs of disease and analysed in week 10.

Statistical analysis. *P* values were calculated using Student's *t*-test. Error bars in graphs represent the s.e.m.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 13 May; accepted 25 June 2008.
Published online 13 August 2008.

1. Taylor, N. A., Van De Ven, W. J. & Creemers, J. W. Curbing activation: proprotein convertases in homeostasis and pathology. *FASEB J.* **17**, 1215–1227 (2003).
2. Dubois, C. M. *et al.* Evidence that furin is an authentic transforming growth factor-beta1-converting enzyme. *Am. J. Pathol.* **158**, 305–316 (2001).
3. Pesu, M., Muul, L., Kanno, Y. & O'Shea, J. J. Proprotein convertase furin is preferentially expressed in T helper 1 cells and regulates interferon gamma. *Blood* **108**, 983–985 (2006).
4. Thomas, G. Furin at the cutting edge: from protein traffic to embryogenesis and disease. *Nature Rev. Mol. Cell Biol.* **3**, 753–766 (2002).
5. Shiryaev, S. A. *et al.* Targeting host cell furin proprotein convertases as a therapeutic strategy against bacterial toxins and viral pathogens. *J. Biol. Chem.* **282**, 20847–20853 (2007).

6. Scamuffa, N. *et al.* Selective inhibition of proprotein convertases represses the metastatic potential of human colorectal tumor cells. *J. Clin. Invest.* **118,** 352–363 (2008).
7. Roebroek, A. J. *et al.* Failure of ventral closure and axial rotation in embryos lacking the proprotein convertase Furin. *Development* **125,** 4863–4876 (1998).
8. Roebroek, A. J. *et al.* Limited redundancy of the proprotein convertase furin in mouse liver. *J. Biol. Chem.* **279,** 53442–53450 (2004).
9. Fink, P. J., Fang, C. A. & Turk, G. L. The induction of peripheral tolerance by the chronic activation and deletion of $CD4^+V\beta5^+$ cells. *J. Immunol.* **152,** 4270–4281 (1994).
10. Tang, Q. & Bluestone, J. A. The $Foxp3^+$ regulatory T cell: a jack of all trades, master of regulation. *Nature Immunol.* **9,** 239–244 (2008).
11. Nakamura, K. *et al.* TGF-β1 plays an important role in the mechanism of $CD4^+CD25^+$ regulatory T cell activity in both humans and mice. *J. Immunol.* **172,** 834–842 (2004).
12. Li, M. O., Wan, Y. Y. & Flavell, R. A. T cell-produced transforming growth factor-beta1 controls T cell tolerance and regulates Th1- and Th17-cell differentiation. *Immunity* **26,** 579–591 (2007).
13. Rubtsov, Y. P. & Rudensky, A. Y. TGFβ signalling in control of T-cell-mediated self-reactivity. *Nature Rev. Immunol.* **7,** 443–453 (2007).
14. Annes, J. P., Munger, J. S. & Rifkin, D. B. Making sense of latent TGFβ activation. *J. Cell Sci.* **116,** 217–224 (2003).
15. Chen, W. *et al.* Conversion of peripheral $CD4^+CD25^-$ naive T cells to $CD4^+CD25^+$ regulatory T cells by TGF-β induction of transcription factor Foxp3. *J. Exp. Med.* **198,** 1875–1886 (2003).
16. Andersson, J. *et al.* $CD4^+Foxp3^+$ regulatory T cells confer infectious tolerance in a TGF-β-dependent manner. *J. Exp. Med.* (in the press).
17. Schon, M. P. *et al.* Mucosal T lymphocyte numbers are selectively reduced in integrin-αE (CD103)-deficient mice. *J. Immunol.* **162,** 6641–6649 (1999).
18. Powrie, F., Carlino, J., Leach, M. W., Mauze, S. & Coffman, R. L. A critical role for transforming growth factor-β but not interleukin 4 in the suppression of T helper type 1-mediated colitis by CD45RB(low) $CD4^+$ T cells. *J. Exp. Med.* **183,** 2669–2674 (1996).
19. Liang, S. *et al.* Conversion of $CD4^+CD25^-$ cells into $CD4^+CD25^+$ regulatory T cells *in vivo* requires B7 costimulation, but not the thymus. *J. Exp. Med.* **201,** 127–137 (2005).
20. Kretschmer, K. *et al.* Inducing and expanding regulatory T cell populations by foreign antigen. *Nature Immunol.* **6,** 1219–1227 (2005).
21. Thornton, A. M. & Shevach, E. M. $CD4^+CD25^+$ immunoregulatory T cells suppress polyclonal T cell activation *in vitro* by inhibiting interleukin 2 production. *J. Exp. Med.* **188,** 287–296 (1998).
22. Travis, M. A. *et al.* Loss of integrin $\alpha_v\beta_8$ on dendritic cells causes autoimmunity and colitis in mice. *Nature* **449,** 361–365 (2007).
23. Bassi, D. E., Fu, J., Lopez de Cicco, R. & Klein-Szanto, A. J. Proprotein convertases: ''master switches'' in the regulation of tumor growth and progression. *Mol. Carcinog.* **44,** 151–161 (2005).
24. Ornatowski, W., Poschet, J. F., Perkett, E., Taylor-Cousar, J. L. & Deretic, V. Elevated furin levels in human cystic fibrosis cells result in hypersusceptibility to exotoxin A-induced cytotoxicity. *J. Clin. Invest.* **117,** 3489–3497 (2007).
25. Sun, C. M. *et al.* Small intestine lamina propria dendritic cells promote *de novo* generation of Foxp3 T reg cells via retinoic acid. *J. Exp. Med.* **204,** 1775–1785 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Bonifacino (National Institute of Child Health and Human Development) and S. Wahl (National Institute of Dental and Craniofacial Research) for providing comments and reading the manuscript, A. Singer (National Cancer Institute) and L. Pobenzinsky (NCI) for providing suggestions, and M. Oukka (Harvard Medical School) for providing Foxp3-GFP mice. This work was supported by the Intramural Research Program of NIAMS, Academy of Finland, the Finnish Medical Foundation, the Maud Kuistila Memorial Foundation, the 'Fonds voor Wetenschappelijk Onderzoek Vlaanderen' and 'Geconcerteerde Onderzoeksactie van de Vlaamse Gemeenschap'.

Author Contributions M.P. designed, performed and interpreted all the experiments and wrote the manuscript. W.T.W. helped to plan, perform and interpret suppression assays as well as some flow cytometry experiments. L.W. performed and interpreted the microarray experiments. L.X. helped to design and perform the TGF-β1 measurements. I.F. interpreted the data and analysed the gut histology. W.S. contributed to the experimental design of conversion and suppression assays and interpreted the data. J.A. helped to plan and perform the *in vitro* conversion assays. E.M.S. contributed to the experimental design of conversion and suppression assays and interpreted the data. M.Q. analysed histopathology. N.B. helped to isolate the gut lymphocytes. A.R. generated and provided the $fur^{f/+}$ animals. Y.B. provided the $RAG2^{-/-}$ and TCR-$\alpha^{-/-}$ animals, contributed to the experimental design of conversion and suppression assays, and interpreted the data. J.C. generated and provided the $fur^{f/+}$ animals and helped to plan and interpret the experiments. J.J.O.'S. oversaw experimental designs, analysed and interpreted all acquired data, and wrote the manuscript.

Author Information The microarray files are deposited in Gene Expression Omnibus under accession number GSE11884. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.P. (pesum@mail.nih.gov).

METHODS

Mice. Mice that expressed floxed *fur* alleles ($fur^{f/f}$)[8] were backcrossed six times with C57BL/6 mice. C57BL/6, CD4-Cre, $RAG2^{-/-}$ and TCR-$\alpha^{-/-}$ mice, all on C57BL/6 background, were from Taconic. CD4-Cre mice were bred with $fur^{f/f}$ animals to generate T-cell-specific furin knockout (CD4cre-$fur^{f/f}$). For *in vitro* conversion assays, CD4cre-$fur^{f/f}$ animals bearing Foxp3-GFP transgene (from M. Oukka) were used. All mice were housed under pathogen-free conditions in accordance with the National Institutes of Health Animal Care and Use Committee.

Flow cytometry. For surface staining, cells were resuspended in fluorescence-activated cell sorting (FACS) staining buffer (PBS + 1% FBS) containing 2 μg ml^{-1} Fc block (BD Pharmingen) and stained on ice for 15–30 min using indicated antibodies from BD Pharmingen or biotinylated anti-latency-associated peptide antibody and goat-IgG from R&D Systems. Intracellular Foxp3 was assessed by intracellular staining according to the manufacturer's instructions, using antibody and buffers from eBiosciences. For intracellular cytokine stainings, cells were activated for 4 h with phorbol myristate acetate and ionomycin; GolgiPlug (BD Pharmingen) was added to cultures after 2 h. Intracellular staining was done with intracellular cytokine staining kit (BD Pharmingen). Cells were analysed using a FACSCalibur or FACSCanto instrument (BD Pharmingen) and data were analysed with Flow-Jo software (Treestar).

Cell purification and culture. For quantification of furin deletion efficiency and induced T_{reg} culture, $CD4^+$ and $CD8^+$ cells were purified by positive selection using magnetic beads (Miltenyi Biotech). Cells were activated for 3 days with plate-bound anti-CD3 and CD28 antibodies (5 μg ml^{-1} in PBS), and further expanded for three additional days in complete 10% FBS RPMI supplemented with IL-2 (50 U ml^{-1}). For Th1 polarization, initial activation of $CD4^+$ cells was done in the presence of IL-12 (10 ng ml^{-1}) and anti-IL-4 antibody (10 μg ml^{-1}). For induced T_{reg} culture, $CD4^+$ cells were activated with plate-bound anti-CD3 (10 μg ml^{-1}) and soluble anti-CD28 (2 μg ml^{-1}) antibodies, IL-2 (100 ng ml^{-1}) and TGF-β1 (5 ng ml^{-1}). For suppression assays and microarray analysis, $CD4^+$ cells were first enriched with a $CD4^+$ T-cell-negative selection kit (Miltenyi) followed by flow cytometry sorting for $CD4^+CD25^-CD45Rb^{hi}$ and $CD4^+CD25^+$ or $CD4^+CD44^{low}CD62L^+$ cells using a MoFlo cell sorter (Dako). For *in vitro* conversion assay, $CD4^+$Foxp3-GFP^+ and $CD4^+$Foxp3-GFP^- cells were purified by flow cytometry.

Western blot and real-time PCR. Western blotting was performed as described[3] using actin (Chemicon), furin (Santa Cruz) or activated TGF-β1 (sc-146, Santa Cruz) antibodies. Total RNA was isolated with RNeasy (Qiagen) and reverse transcribed with complementary DNA synthesis kits (Applied Biosystems). *Fur* exon two specific real-time polymerase chain reaction was performed using an ABI PRISM 7700 Sequence Detection System as described previously[8].

Cytokine and antibody measurements. Mesenteric lymph node T cells (1×10^6 ml^{-1}) were activated for 48 h with plate-bound anti-CD3 (10 μg ml^{-1}) and soluble anti-CD28 (2 μg ml^{-1}; BD Pharmingen). Cytokine levels in supernatants or sera were determined with IL-13 and IL-17 ELISA (R&D Systems) or with mouse Th1/Th2 cytokine or inflammation Cytometric Bead Array kits (BD Pharmingen). Serum-circulating anti-DNA and anti-nuclear antibodies, and IgG1, IgG2 and IgE immunoglobulins, were determined with ELISA kits from Alpha Diagnostic International. To determine TGF-β1 production, FACS-sorted $CD4^+CD25^-$ and $CD4^+CD25^+$ were first stimulated for 48 h with anti-CD3 (10 μg ml^{-1}) and soluble anti-CD28 (2 μg ml^{-1}) in complete RPMI 10% FBS supplemented with IL-2 (50 U ml^{-1}), rested 24 h in complete medium, washed twice with PBS and finally re-stimulated in RPMI + 3% FBS + 1% Nutridoma (Roche) supplemented with IL-2 (50 U ml^{-1}) for another 48 h in the presence or absence of furin inhibitor II (Hexa-d-Arg from Calbiochem). TGF-β1 was measured in the supernatants using a multispecies TGF-β1 ELISA kit (Invitrogen) or ELISA specific for mature/active TGF-β1 (eBiosciences); TGF-β1 concentration in the final stimulation medium was measured and subtracted as background.

Microarray analysis. Total RNA from wild-type and furin-deficient $CD4^+CD44^{low}CD62L^+$ cells was extracted with TRIzol reagent (Invitrogen) according to the manufacturer's instructions. Approximately 500 ng of RNA was labelled using a MessageAmp II Biotin Enhanced kit (Ambion) and hybridized to GeneChip Mouse Genome 430 2.0 arrays (Affymetrix) according to the manufacturers' protocols. Gene expression values were determined using GeneChip Operating Software version 1.1.1. Data were analysed using GeneSpring software GX 7.3.1 (Agilent Technologies). Gene expression was normalized across each chip as well as across all experiments. Genes with an average expression value below 50 and those flagged absent in all samples were deleted from subsequent analysis.

Isolation of gut intraepithelial and lamina propria lymphocytes. To isolate intraepithelial lymphocytes, Peyer's patches were removed and mouse small intestines were cut into segments of 1–3 cm. Gut pieces were placed in RPMI 2% FBS with constant stirring and incubated for 20 min at 37 °C. After incubation, the gut suspension was strained though a sterile strainer, and flow-through was collected in a beaker on ice. To ensure optimum yield of intraepithelial lymphocytes, gut pieces were placed into a 50-ml tube containing 15 ml of serum-free RPMI, shaken vigorously for 30 s, strained as above and further filtered through a 70-μm cell strainer. Cells were sedimented for 10 min at 1,500 r.p.m. at 4 °C. The cell pellet was re-suspended in 30% Percoll and centrifuged at 1,600 r.p.m. at room temperature for 20 min. The supernatant was discarded and the pellet containing intraepithelial lymphocytes was collected for analysis. Isolation of lamina propria lymphocytes was performed as described previously[25].

***In vitro* suppression assay.** Varying numbers of wild-type or furin-deficient $CD4^+CD25^+$ T regulatory cells were cultured in 96-well round-bottom plates with 5×10^4 wild-type effector $CD4^+CD25^-$ T cells along with 5×10^4 $CD90^+$-depleted, irradiated splenocytes used as antigen-presenting cells[21]. Cells were stimulated with 0.5 μg ml^{-1} CD3 antibody (BD Biosciences) for 72 h at 37 °C and 5% CO_2. Cultures were pulsed with [^{3}H]TdR at 1 μCi per well for the last 16 h of culture.

***In vivo* suppression assay.** *In vivo* suppression assays were done as previously described[18]. Briefly, TCR-$\alpha^{-/-}$ mice were injected intravenously with 3×10^5 FACS-sorted wild-type or furin-deficient naive $CD4^+CD25^-CD45Rb^{hi}$ cells with or without 1.5×10^5 wild-type or furin-deficient $CD4^+CD25^+$ T regulatory cells (five mice per group). Mice were monitored weekly for weight loss and signs of disease, and killed in week 10. Total mesenteric lymph node cells were isolated and counted; T-cell numbers and ratios of $CD4^+Foxp3^+$ cells were determined with flow cytometry. For proliferation and IFN-γ production assessment, $RAG2^{-/-}$ mice were reconstituted with carboxyfluorescein succinimidyl ester (CFSE)- labelled naive $CD4^+CD25^-CD45Rb^{hi}$ cells alone or with $CD4^+CD25^+$ T regulatory cells (three mice per group). Congenic markers CD45.1 and CD45.2 were used to distinguish between transferred cells. Mice were analysed on day seven for effector T-cell proliferation and cytokine production by flow cytometry.

***In vitro* conversion assay.** $CD4^+Foxp3^+$ T regulatory cells were activated for 4 days with plate-bound anti-CD3 and IL-2 (100 U ml^{-1}). The activated T_{reg} cells were then co-cultured for an additional 4 days with CFSE-labelled wild-type or furin-deficient $CD4^+Foxp3^-$ responder cells in the presence of splenic dendritic cells at a ratio of 5:5:1 ($CD4^+Foxp3^+$:$CD4^+Foxp3^-$:splenic dendritic cells), anti-CD3 (2 μg ml^{-1}) and IL-2 (100 U ml^{-1}) as indicated. Cytokine-induced conversion of wild-type or furin-deficient $CD4^+Foxp3^-$ effectors was investigated in the absence of T_{reg} cells, but in the presence of dentritic cells and exogenous TGF-β1 (5 ng ml^{-1}). $CFSE^+$ T cells were analysed for Foxp3 expression using flow cytometry.

Statistical analysis. *P* values were calculated using Student's *t*-test; error bars in graphs represent the s.e.m.

LETTERS

Heterochromatin links to centromeric protection by recruiting shugoshin

Yuya Yamagishi[1,2], Takeshi Sakuno[1,3], Mari Shimura[4] & Yoshinori Watanabe[1,2]

The centromere of a chromosome is composed mainly of two domains, a kinetochore assembling core centromere and peri-centromeric heterochromatin regions[1,2]. The crucial role of centromeric heterochromatin is still unknown, because even in simpler unicellular organisms such as the fission yeast *Schizosaccharomyces pombe*, the heterochromatin protein Swi6 (HP1 homologue) has several functions at centromeres, including silencing gene expression and recombination, enriching cohesin, promoting kinetochore assembly, and, ultimately, preventing erroneous microtubule attachment to the kinetochores[1,3–6]. Here we show that the requirement of heterochromatin for mitotic chromosome segregation is largely replaced by forcibly enriching cohesin at centromeres in fission yeast. However, this enrichment of cohesin is not sufficient to replace the meiotic requirement for heterochromatin. We find that the heterochromatin protein Swi6 associates directly with meiosis-specific shugoshin Sgo1, a protector of cohesin at centromeres. A point mutation of Sgo1 (V242E), which abolishes the interaction with Swi6, impairs the centromeric localization and function of Sgo1. The forced centromeric localization of Sgo1 restores proper meiotic chromosome segregation in *swi6Δ* cells. We also show that the direct link between HP1 and shugoshin is conserved in human cells. Taken together, our findings suggest that the recruitment of shugoshin is the important primary role for centromeric heterochromatin in ensuring eukaryotic chromosome segregation.

To delineate the molecular function of centromeric heterochromatin, we explored whether any mutation of *swi6*$^+$ could separate the two major functions of heterochromatin, transcriptional silencing and chromosome segregation. The null allele of *swi6*$^+$, when combined with tagging of Psc3 (a cohesin subunit) with green fluorescent protein (GFP), results in temperature-sensitive growth[5]. This is presumably because the localization of cohesin at the centromere, which is decreased by *swi6Δ*, is further functionally compromised by the tagging of the cohesin subunit. By mutagenizing the *swi6*$^+$ gene on a background of *psc3–GFP*, we were able to isolate a mutant *swi6-sm1* that showed a defect in transcriptional silencing but not in growth at high temperature (Supplementary Fig. 1). We confirmed that the *swi6-sm1* mutation by itself produces defects in the transcriptional silencing of an integrated marker that localizes in the peri-centromeric regions, like *swi6Δ* (Fig. 1a). By contrast, characteristics potentially related to chromosome segregation (sensitivity to thiabendazole, localization of cohesin at centromeres, chromosome lagging at anaphase, and mini-chromosome maintenance) are all intact in *swi6-sm1* cells, as they are in wild-type cells (Fig. 1a–d). These results indicate that the function of transcriptional silencing is separable from chromosome segregation on the Swi6/HP1 protein.

Together with the previous result that cohesin at centromeres is important for chromosome segregation but not for transcriptional silencing[5,6], the foregoing results indicate that the primary requirement of heterochromatin for mitotic chromosome segregation might be the recruitment of cohesin. To explore this assumption, we sought to localize cohesin at the centromere in a heterochromatin-independent way, and to test its ability to suppress the chromosome segregation defects of heterochromatin-negative cells. For this purpose, we endowed Psc3 with its own ability to localize to the centromere by fusing its carboxy-terminal end with two copies of chromodomain (CD), which binds to H3K9me (histone H3 methylated on Lys 9), which locates mainly in the peri-centromeric

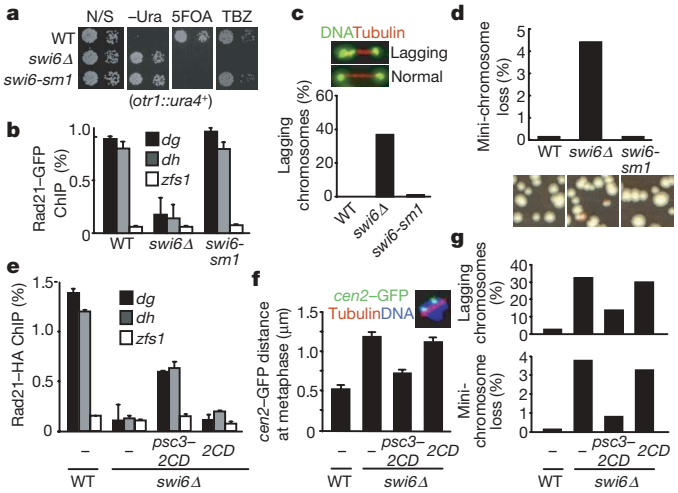


Figure 1 | Forced enrichment of cohesin in the peri-centromeric regions can substitute for the requirement of heterochromatin for mitotic chromosome segregation. **a**, Serial dilutions of the indicated cultures were plated on nonselective (N/S), uracil-lacking (−Ura) or 5-fluoro-orotic acid (5FOA) medium to assay *ura4*$^+$ expression at the outer centromeric repeats. Sensitivity to thiabendazole (TBZ) was assayed similarly. WT, wild type. **b**, The indicated *rad21*$^+$*–GFP* cells were cultured and fixed for ChIP analysis in the peri-centromeric region (*dg* and *dh*) and arm region (*zfs1*). Error bars represent s.d. ($n = 3$). **c**, Frequencies of lagging chromosomes in anaphase cells ($n > 100$) were examined at 18 °C. **d**, Mini-chromosome loss rate per division was assayed in the indicated cells ($n > 1{,}000$). The loss of mini-chromosome results in red colonies as a result of adenine auxotrophy. **e**, Rad21–haemagglutinin (HA) levels were measured in the indicated cells by ChIP analysis. Error bars represent s.d. ($n = 3$). **f**, The indicated cells cultured at 25 °C were shifted to 36 °C for 4 h (metaphase arrest by *cut9-665*), fixed and stained for tubulin. The distance between *cen2*–GFP dots was measured in metaphase (short spindle) cells. Error bars represent s.e.m. ($n = 50$). **g**, Mini-chromosome loss rate per division ($n > 1{,}000$) and frequencies of lagging chromosome at anaphase ($n > 100$) were assayed in the indicated cells.

[1]Laboratory of Chromosome Dynamics, Institute of Molecular and Cellular Biosciences, [2]Graduate Program in Biophysics and Biochemistry, Graduate School of Science, and [3]Promotion of Independence for Young Investigators, University of Tokyo, Yayoi, Tokyo 113-0032, Japan. [4]Department of Intractable Diseases, International Medical Center of Japan, Tokyo 162-8655, Japan.

regions[7]. We confirmed that Psc3–2CD, as well as 2CD, itself localizes at discrete nuclear dots in *swi6Δ* cells, which preserve sufficient H3K9me in the intrinsic heterochromatin regions in spite of the heterochromatin defect, but not in another heterochromatin-defective strain, *clr4Δ*, which lacks H3K9me (ref. 7) (Supplementary Fig. 2a). As predicted, the additional expression of Psc3–2CD (but not of 2CD alone) improved the localization of the cohesin complex to the peri-centromeric regions and also centromeric cohesion in *swi6Δ* cells (Fig. 1e, f). We confirmed that the expression of Psc3–2CD does not restore transcriptional silencing in *swi6Δ* cells (Supplementary Fig. 3a). To explore the impact of cohesion recovery on the defective chromosome segregation in *swi6Δ* cells, we used a mini-chromosome loss assay. The fidelity of chromosome segregation was largely restored by the expression of Psc3–2CD but not that of 2CD (Fig. 1g). Consistently, the frequency of lagging chromosomes in anaphase reduced in *swi6Δ Psc3–2CD* cells but not in *swi6Δ 2CD* cells (Fig. 1g). These results unequivocally validate the previously expected, but not proven, notion that the primary requirement of heterochromatin for mitotic chromosome segregation is the enrichment of cohesin at the centromere[5,6].

In meiotic chromosome segregation, monopolar attachment of sister kinetochores to the spindle is established in metaphase I; sister chromatids therefore move together to the same side of the zygote (reductional division) in the following anaphase I. During meiosis I, meiosis-specific shugoshin Sgo1 and its partner, protein phosphatase 2A (PP2A), protect the centromeric cohesin from separase cleavage. Bipolar attachment at the following meiosis II is therefore secured by the residual cohesion at the centromere[8–11] (Fig. 2a). As with *sgo1Δ* cells, *swi6Δ* cells undergo intact meiosis I but suffer a nondisjunction of sister chromatids in meiosis II (refs 8, 12) (Fig. 2a, b). This can be explained by the fact that meiotic prophase *swi6Δ* cells have a decrease in the primary enrichment of cohesin (including Rec8–Psc3) to 30–40% in peri-centromeric regions[12], which should be protected by Sgo1 in the following anaphase I. The transcriptional silencing of Swi6 is not relevant to this function, because *swi6-sm1* cells have intact meiotic chromosome segregation (Supplementary Fig. 4). We therefore assumed that the forced localization of cohesin at the centromere would also restore the meiotic defects of *swi6Δ* cells. However, the expression of Psc3–2CD scarcely restored the *swi6Δ* defect in meiosis II (data not shown; see also Fig. 2g). This result indicates that the primary requirement of heterochromatin for meiotic chromosome segregation may be something other than cohesin enrichment.

In a search, involving a yeast two-hybrid assay, for proteins that interact with Sgo1, we frequently isolated Swi6. This interaction was confirmed by co-immunoprecipitation in extracts of fission yeast (Fig. 2c). Consistent with the fact that these proteins become enriched in the peri-centromeric regions[8], Sgo1 and Swi6 largely co-localized in the cells at metaphase I (Fig. 2d). These results, together with the above assumption, prompted us to link Swi6 to the Sgo1 function. Accordingly, the measurement of fluorescence intensity of centromeric Sgo1–GFP in metaphase I-arrested cells revealed that Sgo1 localization is impaired in *swi6Δ* cells (Fig. 2e), although meiotic expression of Sgo1 protein was not affected by *swi6Δ* (Supplementary Fig. 5). These results indicate that Swi6 is crucial in localizing Sgo1 and thereby promotes the protection of cohesin from separase during anaphase I. To determine whether Sgo1 localization is the primary meiotic target of Swi6, we fused Sgo1 with CD to localize Sgo1 at centromeres in a heterochromatin-independent way and examined its ability to suppress meiotic defects in *swi6Δ* cells. Sgo1–CD did indeed localize at the centromere regardless of *swi6Δ* (Fig. 2f). The expression of Sgo1–CD in place of endogenous Sgo1 mostly suppressed nondisjunction during meiosis II in *swi6Δ* cells (Fig. 2g; compare the *2CD sgo1*$^{+}$ and *2CD sgo1–CD* columns). This suppression requires the amino-terminal coiled-coil domain of Sgo1, a region that associates with PP2A (see below), which is indicative of the specificity of the suppression (Supplementary Fig. 6). Moreover, the suppression became nearly complete when Psc3–2CD was co-expressed (Fig. 2g; compare the rightmost two columns). These results indicate strongly that the centromeric recruitment of Sgo1 is the primary role for heterochromatin in meiotic chromosome segregation, although the enrichment of cohesin also contributes to ensure that sufficient cohesin is protected by Sgo1 until meiosis II. The defect in disjunction at meiosis II is partial in *swi6Δ* cells in comparison with *sgo1Δ* cells (Fig. 2b), and residual centromeric signals of Sgo1 are detectable in *swi6Δ* cells (Fig. 2e), indicating the possible existence of an additional Swi6-independent pathway for centromeric Sgo1 localization (also see Supplementary Fig. 7).

To delineate the interaction of Swi6 and Sgo1, we made several truncations of these proteins and examined their interaction by two-hybrid assay. We found that the interaction depends on the chromoshadow domain (CSD) of Swi6 and a conserved amino-acid residue Phe 324 within it that is known to mediate an interaction with a specific pentapeptide sequence (P/L)xVx(M/I/L/V)[13] (Fig. 3a). In

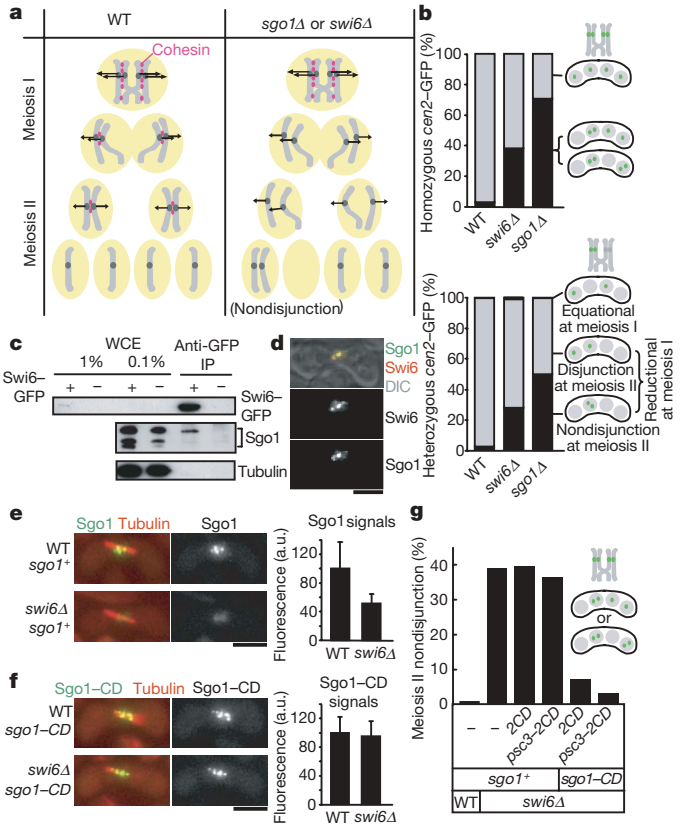


Figure 2 | Requirement of heterochromatin protein Swi6 for meiosis is replaced by forced localization of Sgo1 at the centromere. **a**, Schematic drawing of the behaviour of homologous chromosomes during meiosis. The location of cohesin (red oval) is indicated. WT, wild type. **b**, Both (top) or one (bottom) of the homologues marked with *cen2*–GFP were monitored for segregation during meiosis in the indicated zygotes ($n > 100$). **c**, Whole-cell extracts (WCE) were prepared from proliferating *swi6*$^{+}$*–GFP* or non-tagged cells ectopically expressing Sgo1 and immunoprecipitated (IP) with anti-GFP antibody. The precipitates were examined by western blotting with anti-Sgo1 antibody. The lower bands of Sgo1 represent the degradation products. **d**, Swi6–tdTomato and Sgo1–GFP expressed from endogenous promoters were detected in meiosis I. **e**, Sgo1–GFP signals at metaphase I were measured and compared between wild-type and *swi6Δ* cells. The spindles were visualized by expressing mCherry–Atb2. Error bars represent s.d. ($n = 30$). **f**, Sgo1–GFP–CD signals were measured and compared between wild-type and *swi6Δ* cells. Error bars represent s.d. ($n = 30$). Scale bars, 5 μm. **g**, Both homologues marked with *cen2*–GFP in the indicated zygotes were monitored for segregation during meiosis ($n > 100$). The frequency of zygotes undergoing nondisjunction in either or both meiosis II divisions is shown.

contrast, the Sgo1 peptide responsible for the interaction was limited to 26 residues (222–247), including VCVCI (240–244), which is similar to the CSD-binding motif (Fig. 3b). Accordingly, the replacement of Val 242 with Glu (VE) in Sgo1 abolished the interaction with Swi6 while preserving the interaction with Par1, a subunit of PP2A. An immunoprecipitation assay also supports the loss of the interaction of Sgo1-VE with Swi6 (Fig. 3c). To explore the impact of the loss of the interaction *in vivo*, we expressed *sgo1-VE* from the endogenous promoter and examined the localization of Sgo1. Immunofluorescence assays indicate that the centromeric localization of Sgo1-VE is significantly decreased in comparison with wild-type Sgo1 (Fig. 3d). To delineate the loss of localization more precisely, we performed a chromatin immunoprecipitation (ChIP) assay. The results indicate that Sgo1-VE largely loses the ability to localize in the peri-centromeric region, although Swi6 localization is invariant in this mutant (Fig. 3e). These data indicate that the interaction between Swi6 and Sgo1 is important for Sgo1 localization at centromeres and that Sgo1 acts downstream of heterochromatin assembly at centromeres. The assay of chromosome segregation further revealed that *sgo1-VE* cells provoke nondisjunction in meiosis II, similarly to *swi6Δ* cells (Fig. 3f). The Sgo1-VE protein, when fused with CD and thereby localized to the centromere, can perform its full function in protecting Rec8 (Fig. 3f and Supplementary Fig. 8). These results indicate that Sgo1-VE is solely deficient in the ability to localize at centromeres. Likewise, *sgo1-VE–CD* suppresses *swi6Δ* defects to a similar extent to that of wild-type *sgo1–CD* (Fig. 3f). We noticed that a slightly greater defect is observed in *swi6Δ* cells than in *sgo1-VE* cells, and this is also true in *swi6Δ sgo1-VE–CD* cells compared with *sgo1-VE–CD* cells. This can be accounted for by the difference in the primary localization of cohesin, which is decreased in *swi6Δ* cells[12] but is intact even in *sgo1Δ* cells[9]. Taken together, these results highlight the importance of Val 242 in Sgo1 for Swi6-mediated centromeric localization and the requirement of this interaction for the protection of centromeric cohesin throughout meiosis I division.

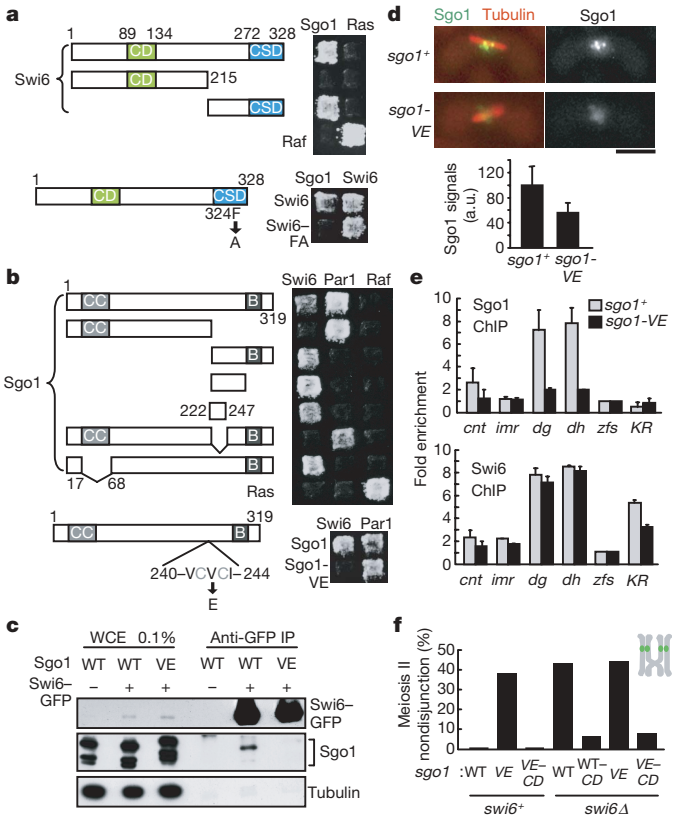


Figure 3 | Association of Swi6 with Sgo1 is required for the centromeric localization and function of Sgo1. **a**, Yeast two-hybrid assay indicates that Swi6 interacts with Sgo1 through the CSD. The pair of Ras and Raf acts as positive control. The point mutation F324A (FA) in the CSD abolishes the interaction with Sgo1, while preserving the ability to dimerize. The transformants were grown on plates lacking histidine. **b**, Yeast two-hybrid assay with Sgo1 deletions. Sgo1 interacts with Par1 through the coiled-coil region (CC), and the interaction with Swi6 takes place through the upstream region (222–247 residues) of the conserved basic region (B). A point mutation, V242E (VE), in this region abolishes the interaction. **c**, Whole-cell extracts (WCE) were prepared from mitotic cells ectopically expressing wild-type (WT) Sgo1 or Sgo1-VE protein. Swi6–GFP was immunoprecipitated (IP) with anti-GFP antibody to test for precipitation together with Sgo1. **d**, Sgo1–GFP and Sgo1-VE–GFP were detected at metaphase I and their signals were measured. Error bars represent s.d. ($n = 30$). Scale bar, 5 μm. **e**, A ChIP assay was used to measure Sgo1 (top) and Swi6 (bottom) levels throughout the indicated chromosomal sites (*cnt* and *imr* locate at the core centromere and *KR* at the silent mating-type loci) in cells arrested at metaphase I. Fold enrichment to the arm region (*zfs*) is shown. Error bars represent s.d. ($n = 2$). **f**, Both homologues marked with *cen2*–GFP in the indicated cells were monitored for segregation during meiosis ($n > 100$). The frequency of zygotes undergoing nondisjunction in either or both meiosis II divisions is shown.

Although the protective function of shugoshin is limited in meiosis in yeast (no shugoshin has a function in cohesin protection in mitosis[14]), metazoan shugoshin protects cohesin from the prophase dissociation pathway[15–17], which is accompanied by extensive chromosome condensation and the resolution of chromosomes during mitosis. We then examined whether the link between heterochromatin protein HP1 and shugoshin-dependent centromeric protection is evolutionarily conserved. In mammalian cells, HP1α is the major isoform of HP1; it localizes in the peri-centromeric region[18,19]. Immunoprecipitation with a chromatin extract from mitotic 293T cells indicates that hSgo1 precipitates together with HP1α as well as PP2A (Fig. 4a). Moreover, two-hybrid assays suggest that hSgo1 potentially interacts with all HP1 isoforms through the CSD (Supplementary Fig. 9) and that, at least, the interaction of HP1α is mediated by the conserved residue Trp 174 within the CSD (Fig. 4b). In contrast, the deletions of hSgo1 suggest that the amino-acid residues required for the interaction are located between positions 451 and 456 (PVVKIR), which contains a putative CSD binding motif[13], and the replacement of Val 453 with Glu (hSgo1-VE) results in the loss of the interaction with HP1α (Fig. 4c). Thus, the manner of interaction between HP1α and hSgo1 is highly conserved with that of Swi6 and Sgo1 in fission yeast.

In mammals, HP1 associating at centromeric heterochromatin is largely displaced at mitosis, but a small population remains at centromeres[18,20,21] (Fig. 4d). Immunofluorescence of enhanced GFP (EGFP)-tagged HP1α expressed in HeLa cells clearly gives a signal at the inner centromere[22], co-localizing with hSgo1 (Supplementary Fig. 10). To examine the requirement of HP1 for hSgo1 localization, we treated HeLa cells with short interfering RNA (siRNA) against HP1α. Cell proliferation was obviously delayed in the HP1α siRNA-treated cells, although the mitotic defect was not obvious in unperturbed mitosis (data not shown). However, if the cells were arrested at prometaphase by the addition of nocodazole, hSgo1 localization was abolished in a subpopulation of HP1α siRNA-treated cells but not in control cells (Fig. 4e). Accordingly, the centromeric cohesion was largely dissociated in the hSgo1-lacking chromosomes but not in chromosomes containing hSgo1 (Fig. 4f, magnified panels in the HP1α siRNA-treated spread). The imperfect penetrance of the phenotype can be accounted for in part by the difference in duration of mitotic arrest between cells; spreads with less condensed chromosomes all showed normal localization of hSgo1 even in HP1α siRNA-treated cells (Fig. 4f). Although previous assays in vertebrate cells suggested that the heterochromatin structure at the centromere is required for cohesin retention at metaphase[23,24], our current results suggest that this may be mediated through shugoshin localization. Considering the report that the H3K9me pathway is dispensable for centromeric protection in mouse fibroblasts[25], HP1α responsible for hSgo1 localization might use the hinge domain of HP1α for

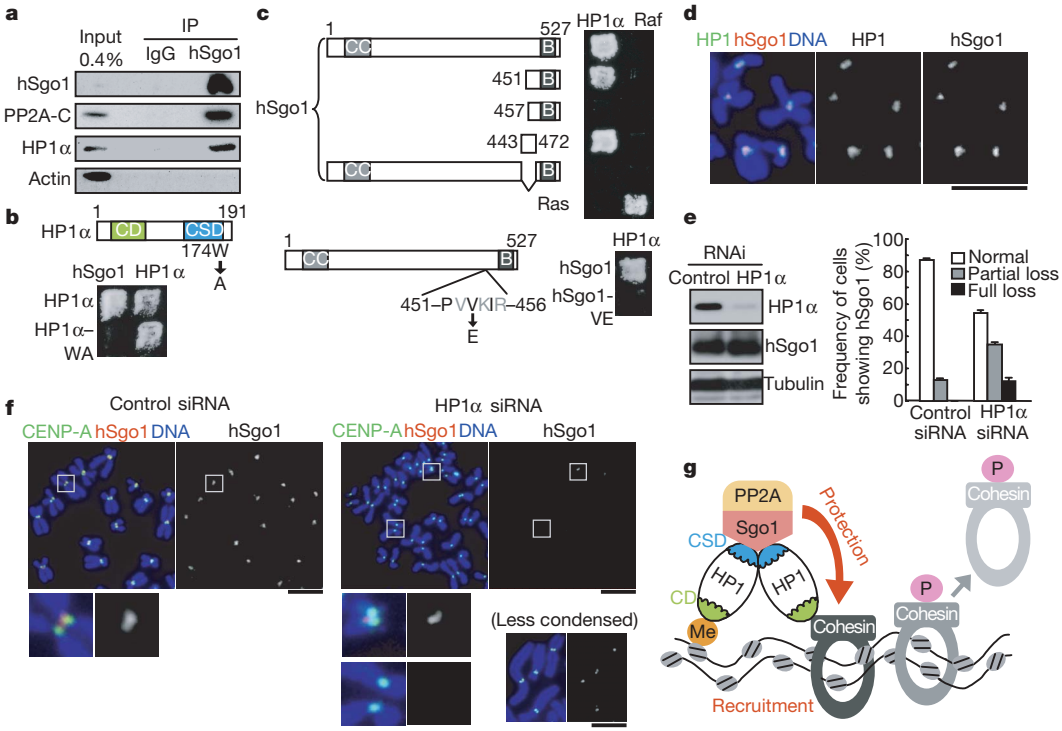


Figure 4 | The human heterochromatin protein HP1α interacts with human Sgo1 (hSgo1) and is required for its maintenance at centromeres in mitotic chromosomes. a, A chromatin extract of nocodazole-arrested 293T cells was immunoprecipitated (IP) and analysed by western blotting. **b**, A yeast two-hybrid assay indicating that HP1α interacts with hSgo1 through Trp 174 within the CSD. **c**, Yeast two-hybrid assay using hSgo1 deletions and point mutation. **d**, Spread chromosomes of HeLa cells were immunostained for hSgo1, HP1 and DNA. **e**, HeLa cells treated with control or HP1α siRNA were cultured for 48 h and for a further 4 h after the addition of nocodazole. Depletion of HP1α was examined by western blotting. Spread chromosomes were immunostained for CENP-A, hSgo1 and DNA. Spreads ($n > 50$) were classified according to hSgo1 staining patterns into normal, partial loss (more than 30% hSgo1-negative chromosomes per cell) or full loss. Error bars represent s.d. ($n = 3$). **f**, Representative spreads assayed in **e** are shown magnified in the bottom panel. A spread of an HP1α siRNA-treated cell with less-condensed chromosomes is also shown at the bottom right. Scale bars, 5 μm. **g**, A schematic model illustrating how Swi6/HP1 protects centromeric cohesion.

centromeric retention[19] rather than the CD–H3K9me interaction that is disrupted at mitosis[20,21]. Taken together, our data suggest that HP1 has a pivotal function in localizing and/or maintaining hSgo1 at centromeres through their direct association in mammalian cells as in fission yeast (Fig. 4g).

Accumulating evidence suggests that the heterochromatin protein Swi6/HP1 functions in recruiting several cellular 'effector' proteins to a specific chromatin site and thereby in regulating various chromosomal processes[3]. Given that almost all eukaryotic chromosomes carry heterochromatin in the peri-centromeric regions, Swi6/HP1 may constitute a crucial basis for centromere function. Indeed, recent studies suggest that heterochromatin influences the process of kinetochore establishment at the centromere[4,26,27]. Here we have addressed the primary requirement of heterochromatin for mitotic and meiotic chromosome segregation in a simple unicellular organism, fission yeast. Although the enrichment of cohesin in peri-centromeric regions is an important requirement of heterochromatin in ensuring mitotic chromosome segregation, our analysis reveals that the most crucial requirement during meiosis is the recruitment and/or maintenance of shugoshin at centromeres. Moreover, our analysis demonstrates that the mechanism by which the heterochromatin protein promotes shugoshin-mediated centromeric protection is conserved in human cells. Thus, shugoshin recruitment might be a hitherto unknown primary role for centromeric heterochromatin in eukaryotic chromosomes.

METHODS SUMMARY

All *Schizosaccharomyces pombe* strains used in this study are listed in Supplementary Table 1. Deletion and tagging of endogenous *sgo1*$^+$, *swi6*$^+$, *rec8*$^+$ and *rad21*$^+$ by GFP, mCherry or tdTomato were performed with the PCR-based gene targeting method for *S. pombe*. To quantify the centromeric fluorescent signals, in-focus images of Sgo1–GFP cells were taken with MetaMorph imaging software (Universal Imaging). We measured the maximum intensity of centromeric signals within cells and subtracted the average of background intensity in the nuclei. A ChIP assay was conducted as described previously[8]. A mini-chromosome loss assay was performed as described previously[5], by culturing cells carrying mini-chromosome Ch16 in medium lacking adenine and plating them onto adenine-limiting plates at 30 °C. Immunoprecipitation was conducted essentially as described previously[14]. Immunofluorescence staining of the spread chromosome of HeLa cells was performed as described[28], using primary antibodies against hSgo1 (1:1,000 dilution) and HP1 (1:100 dilution; Santa Cruz Biotechnology, Inc.) or CENP-A (1:100 dilution; Abcam), which were diluted with buffer A containing 10% goat serum or 0.1% BSA, and secondary antibodies: Alexa Fluor 488 anti-rabbit IgG (1:400 dilution; Molecular Probes, Inc.) and Alexa Fluor 546 anti-rabbit IgG (1:400 dilution; Molecular Probes, Inc.), or Alexa Fluor 546 anti-goat IgG (1:1,000 dilution; Molecular Probes, Inc.) and Alexa Fluor 488 anti-rabbit IgG (1:400 dilution; Molecular Probes, Inc.).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 8 April; accepted 27 June 2008.
Published online 17 August 2008.

1. Pidoux, A. L. & Allshire, R. C. The role of heterochromatin in centromere function. *Phil. Trans. R. Soc. B* **360**, 569–579 (2005).
2. Amor, D. J., Kalitsis, P., Sumer, H. & Choo, K. H. A. Building the centromere: from foundation proteins to 3D organization. *Trends Cell Biol.* **14**, 359–368 (2004).
3. Grewal, S. I. & Jia, S. Heterochromatin revisited. *Nature Rev. Genet.* **8**, 35–46 (2007).
4. Folco, H. D., Pidoux, A. L., Urano, T. & Allshire, R. C. Heterochromatin and RNAi are required to establish CENP-A chromatin at centromeres. *Science* **319**, 94–97 (2008).
5. Nonaka, N. *et al.* Recruitment of cohesin to heterochromatic regions by Swi6/HP1 in fission yeast. *Nature Cell Biol.* **4**, 89–93 (2002).
6. Bernard, P. *et al.* Requirement of heterochromatin for cohesion at centromeres. *Science* **294**, 2539–2542 (2001).

7. Nakayama, J., Rice, J. C., Strahl, B. D., Allis, C. D. & Grewal, S. I. Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* **292,** 110–113 (2001).
8. Kitajima, T. S., Kawashima, S. A. & Watanabe, Y. The conserved kinetochore protein shugoshin protects centromeric cohesion during meiosis. *Nature* **427,** 510–517 (2004).
9. Rabitsch, K. P. *et al.* Two fission yeast homologs of *Drosophila* Mei-S332 are required for chromosome segregation during meiosis I and II. *Curr. Biol.* **14,** 287–301 (2004).
10. Kitajima, T. S. *et al.* Shugoshin collaborates with protein phosphatase 2A to protect cohesin. *Nature* **441,** 46–52 (2006).
11. Riedel, C. G. *et al.* Protein phosphatase 2A protects centromeric sister chromatid cohesion during meiosis I. *Nature* **441,** 53–61 (2006).
12. Kitajima, T. S., Yokobayashi, S., Yamamoto, M. & Watanabe, Y. Distinct cohesin complexes organize meiotic chromosome domains. *Science* **300,** 1152–1155 (2003).
13. Smothers, J. F. & Henikoff, S. The HP1 chromo shadow domain binds a consensus peptide pentamer. *Curr. Biol.* **10,** 27–30 (2000).
14. Kawashima, S. A. *et al.* Shugoshin enables tension-generating attachment of kinetochores by loading Aurora to centromeres. *Genes Dev.* **21,** 420–435 (2007).
15. Salic, A., Waters, J. C. & Mitchison, T. J. Vertebrate shugoshin links sister centromere cohesion and kinetochore microtubule stability in mitosis. *Cell* **118,** 567–578 (2004).
16. McGuinness, B. E., Hirota, T., Kudo, N. R., Peters, J.-M. & Nasmyth, K. Shugoshin prevents dissociation of cohesin from centromeres during mitosis in vertebrate cells. *PLoS Biol.* **3,** e86 (2005).
17. Kitajima, T. S., Hauf, S., Ohsugi, M., Yamamoto, T. & Watanabe, Y. Human Bub1 defines the persistent cohesion site along the mitotic chromosome by affecting shugoshin localization. *Curr. Biol.* **15,** 353–359 (2005).
18. Minc, E., Allory, Y., Worman, H. J., Courvalin, J. C. & Buendia, B. Localization and phosphorylation of HP1 proteins during the cell cycle in mammalian cells. *Chromosoma* **108,** 220–234 (1999).
19. Hayakawa, T., Haraguchi, T., Masumoto, H. & Hiraoka, Y. Cell cycle behavior of human HP1 subtypes: distinct molecular domains of HP1 are required for their centromeric localization during interphase and metaphase. *J. Cell Sci.* **116,** 3327–3338 (2003).
20. Fischle, W. *et al.* Regulation of HP1-chromatin binding by histone H3 methylation and phosphorylation. *Nature* **438,** 1116–1122 (2005).
21. Hirota, T., Lipp, J. J., Toh, B. H. & Peters, J. M. Histone H3 serine 10 phosphorylation by Aurora B causes HP1 dissociation from heterochromatin. *Nature* **438,** 1176–1180 (2005).
22. Sugimoto, K., Tasaka, H. & Dotsu, M. Molecular behavior in living mitotic cells of human centromere heterochromatin protein HP1α ectopically expressed as a fusion to red fluorescent protein. *Cell Struct. Funct.* **26,** 705–718 (2001).
23. Fukagawa, T. *et al.* Dicer is essential for formation of the heterochromatin structure in vertebrate cells. *Nature Cell Biol.* **6,** 784–791 (2004).
24. Guenatri, M., Bailly, D., Maison, C. & Almouzni, G. Mouse centric and pericentric satellite repeats form distinct functional heterochromatin. *J. Cell Biol.* **166,** 493–505 (2004).
25. Koch, B., Kueng, S., Ruckenbauer, C., Wendt, K. S. & Peters, J. M. The Suv39h-HP1 histone methylation pathway is dispensable for enrichment and protection of cohesin at centromeres in mammalian cells. *Chromosoma* **117,** 199–210 (2008).
26. Obuse, C. *et al.* A conserved Mis12 centromere complex is linked to heterochromatic HP1 and outer kinetochore protein Zwint-1. *Nature Cell Biol.* **6,** 1135–1141 (2004).
27. Okada, T. *et al.* CENP-B controls centromere formation depending on the chromatin context. *Cell* **131,** 1287–1300 (2007).
28. Toyoda, Y. & Yanagida, M. Coordinated requirements of human topo II and cohesin for metaphase centromere alignment under Mad2-dependent spindle checkpoint surveillance. *Mol. Biol. Cell* **17,** 2287–2302 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Hauf for reading the manuscript critically; R. Allshire, S. Grewal and the Yeast Genetic Resource Center (YGRC) for yeast strains; J.-i. Nakayama for Swi6 antibody; and all the members of our laboratory for their valuable support and discussion. This work was supported in part by the Global COE programme (Integrative life Science Based on the Study of Biosignaling Mechanisms), MEXT, Japan, the Toray Science Foundation (to Y.W.), Special Coordination Funds for Promoting Science and Technology (to T.S.) and Grants-in-Aid for Research on Advanced Medical Technology, Ministry of Health, Labour and Welfare (to M.S.) and for Specially Promoted Research, MEXT, Japan (to Y.W.).

Author Contributions Experiments in Fig. 1 were performed mainly by T.S., those in Figs 2, 3 and 4a–c by Y.Y., and those in Fig. 4d–f by M.S. Experimental design and interpretation of data were conducted by all authors. Y.W. planned the project and wrote the paper with input from co-authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to Y.W. (ywatanab@iam.u-tokyo.ac.jp).

METHODS

***Schizosaccharomyces pombe* strains.** *sgo1*$^{+}$*–flag–GFP* was created as described previously[8]. *sgo1*$^{+}$*–flag–GFP–CD* was created by inserting the CD of Swi6 (residues 69–143) into the C terminus of *sgo1*$^{+}$*–flag–GFP* at an endogenous *sgo1*$^{+}$ locus using a PCR-based gene targeting method. We abbreviate Sgo1–Flag–GFP as Sgo1–GFP or simply Sgo1, and Sgo1–Flag–GFP–CD as Sgo1–GFP–CD or Sgo1–CD. To express Psc3–CFP–2CD, a sequence encoding cyan fluorescent protein (CFP) and two copies of the CD of Chp1 (residues 1–97) were fused to the C terminus of Psc3 and cloned under the promoter P*adh41* (a weak version of the *adh1*$^{+}$ promoter). The resulting plasmid was linearized and integrated at the *lys1*$^{+}$ locus of chromosome I by using the *nat*^{r} marker; Psc3–CFP–2CD was thereby expressed in addition to the endogenous Psc3. To express Sgo1(CCΔ)–Flag–GFP–CD, a sequence of Sgo1–Flag–GFP–CD with the endogenous promoter of *sgo1*$^{+}$ was cloned. Then the sequence corresponding to residues 18–67 of Sgo1 was deleted by using a PCR-based method. The resulting plasmid was linearized and integrated at the locus adjacent to the *zfs1*$^{+}$ gene of chromosome II (we refer to this as the *z* locus) using the *hyg*^{r} marker. To express mCherry–Atb2, a sequence encoding mCherry was fused to the N terminus of *atb2*$^{+}$, cloned under the promoter P*adh15* (a weak version of the *adh1*$^{+}$ promoter), and integrated into the *z* locus. To overexpress Sgo1 in mitotic cells, the endogenous *sgo1*$^{+}$ promoter was replaced with the *nmt1*$^{+}$ promoter or Sgo1 was expressed by the pREP1(*nmt1*$^{+}$ promoter)-*sgo1*$^{+}$ plasmid. To decrease anaphase-promoting complex activity during meiosis, and thereby arrest meiosis at metaphase I, we replaced the promoters of *slp1*$^{+}$ and *cut23*$^{+}$ with that of *rad21*$^{+}$, which is repressed during meiosis[8].

Two-hybrid assay. The constructs of *sgo1*, *swi6* and *HP1* derivatives were amplified by PCR, cloned into pBTM116 vectors and used as bait. The constructs of *swi6*, *par1*, *HP1* and *hSGO1* derivatives were amplified by PCR, cloned into pActII, pVP16 or pGADT7 vectors and used as prey. These plasmids were transformed into the L40 strain of *Saccharomyces cerevisiae*. Plates lacking histidine were used as a selective medium.

Immunoprecipitation from fission yeast extracts. Cultured cells were harvested, washed with HB buffer (25 mM MOPS pH 7.2, 15 mM $MgCl_2$, 15 mM EGTA, 60 mM β-glycerophosphate, 0.1 mM sodium orthovanadate, 0.1 mM NaF, 15 mM *p*-nitrophenylphosphate, 1% Triton-X100, 1 mM dithiothreitol, 1 mM phenylmethylsulphonyl fluoride and Complete protease inhibitor (Roche)) and disrupted with a Multi-bead shocker (Yasui Kikai); the supernatants were collected after centrifugation. Cell extracts were incubated with anti-GFP polyclonal antibodies (Living Colours Full-length A.v. Polyclonal Antibody; BD Biosciences) for 1 h at 4 °C. Protein A beads (Amersham) were added and incubation was continued for 2 h at 4 °C. After washing the precipitates with HB buffer, we analysed them by SDS–PAGE and western blotting with anti-GFP (1:1,000 dilution; Roche), anti-Sgo1 (ref. 8) (1:1,000 dilution) and TAT1 (1:5,000 dilution) antibodies.

ChIP assay. The anti-GFP polyclonal antibody (BD Biosciences) and anti-Swi6 polyclonal antibody were used for immunoprecipitation. DNA prepared from whole-cell extracts or immunoprecipitated fractions was analysed by quantitative PCR with the ABI PRISM 7000 system (Applied Biosystems) with SYBR Premix Ex Taq (Perfect Real Time; Takara). The primers used for PCR were all described previously[8,14]. We included an untagged strain to account for non-specific binding in the ChIP fractions. The immunoprecipitation ratios in each region were divided by that of the *zfs1* region (arm region), giving the enrichment scores.

Immunoprecipitation from human cell extracts. 293T cells were harvested after treatment with 330 nM nocodazole for 12 h and were lysed by freezing and thawing. Insoluble materials were collected and chromatin proteins were solubilized by treatment with micrococcal nuclease. The cleared chromatin extract was incubated for 3 h at 4 °C with anti-hSgo1 or rabbit IgG coupled to protein A beads. After washing the beads, we analysed the immunoprecipitates by western blotting with anti-hSgo1 (ref. 17) (1:1,000 dilution), anti-PP2A-C (1:1,000 dilution; BD Biosciences), anti-HP1α (1:1,000 dilution; Chemicon) and anti-actin (1:1,000 dilution; Santa Cruz Biotechnology, Inc).

Treatment of HeLa cells with siRNA and immunofluorescence labelling of spread chromosomes. siRNA (200 nM) was used for transfection. The sequences of siRNA for HP1α were previously reported[26] (we obtained essentially the same result with a different siRNA, namely 5′-GGGAGAAGUCA GAAAGUAATT-3′). We treated the siRNA 5′-UAAGGCUAUGAAGAG AUAC-3′ (siCONTROL Non-targeting siRNA no. 2; Dharmacon) as control siRNA. After treatment with siRNAs, HeLa cells were cultured for 48 h and treated with nocodazole (0.1 μg ml^{-1}) for 4 h. The rounded-up cells at mitotic phase were collected and gently resuspended in 75 mM KCl and incubated for 10 min at room temperature (20–25 °C). These hypotonically treated cells were sedimented onto slides by cytocentrifugation (Cytospin 4; ThermoShandon) at 600 r.p.m. (41*g*) for 5 min. The area around the spreads was marked with PAP-PEN (Zymed Laboratories, Inc.) and washed with buffer A (10 mM Tris-HCl pH 8.0, 120 mM NaCl, 0.5 mM EDTA, 0.1% Triton X-100) for at least 10 min. The slides were then incubated for 1 h at room temperature in a humid chamber with primary antibodies. After gentle washing three times with buffer A (once for 1 min, twice for 5 min), the slides were further incubated with secondary antibodies for 30 min at room temperature in a humid chamber, and gently washed again three times with buffer A. The cells were then fixed with 4% paraformaldehyde in buffer A for 15 min at room temperature. After gentle washing in PBS-EGTA and staining with 4,6-diamidino-2-phenylindole, the slides were mounted in Vectashield for immunofluorescence microscopy.

Immunostaining of HeLa cells expressing EGFP-tagged HP1α. Immunofluorescence staining of human cells was performed as described[29]. The sequence of HP1α was cloned into the pEGFP-C1 vector (Clontech), and the resulting pEGFP-C1-HP1α plasmid was transfected into HeLa cells with the use of FuGENE6 reagent (Roche). HeLa cells expressing EGFP–HP1α were spun onto glass slides with a Cytospin cytocentrifuge (Thermo Electron Corporation) and immunostained with anti-hSgo1 (1:1,000 dilution), anti-GFP (1:1,000 dilution; Molecular Probes), and anti-centromere antibody (1:100 dilution; MBL). Secondary antibodies were Alexa Fluor 488 anti-rabbit antibodies (1:1,000 dilution; Molecular Probes), Alexa Fluor 568 anti-mouse antibodies (1:1,000 dilution; Molecular Probes) and Alexa Fluor 647 anti-human antibodies (1:1,000 dilution; Molecular Probes). Images were captured by DeltaVision SoftWorx software (Applied Precision) and processed by deconvolution and *Z*-stack projection.

29. Lee, J. *et al.* Unified mode of centromeric protection by shugoshin in mammalian oocytes and somatic cells. *Nature Cell Biol.* **10**, 42–52 (2008).

CORRIGENDUM

doi:10.1038/nature07253

Genome analysis of the platypus reveals unique signatures of evolution

Wesley C. Warren, LaDeana W. Hillier, Jennifer A. Marshall Graves, Ewan Birney, Chris P. Ponting, Frank Grützner, Katherine Belov, Webb Miller, Laura Clarke, Asif T. Chinwalla, Shiaw-Pyng Yang, Andreas Heger, Devin P. Locke, Pat Miethke, Paul D. Waters, Frédéric Veyrunes, Lucinda Fulton, Bob Fulton, Tina Graves, John Wallis, Xose S. Puente, Carlos López-Otín, Gonzalo R. Ordóñez, Evan E. Eichler, Lin Chen, Ze Cheng, Janine E. Deakin, Amber Alsop, Katherine Thompson, Patrick Kirby, Anthony T. Papenfuss, Matthew J. Wakefield, Tsviya Olender, Doron Lancet, Gavin A. Huttley, Arian F. A. Smit, Andrew Pask, Peter Temple-Smith, Mark A. Batzer, Jerilyn A. Walker, Miriam K. Konkel, Robert S. Harris, Camilla M. Whittington, Emily S. W. Wong, Neil J. Gemmell, Emmanuel Buschiazzo, Iris M. Vargas Jentzsch, Angelika Merkel, Juergen Schmitz, Anja Zemann, Gennady Churakov, Jan Ole Kriegs, Juergen Brosius, Elizabeth P. Murchison, Ravi Sachidanandam, Carly Smith, Gregory J. Hannon, Enkhjargal Tsend-Ayush, Daniel McMillan, Rosalind Attenborough, Willem Rens, Malcolm Ferguson-Smith, Christophe M. Lefèvre, Julie A. Sharp, Kevin R. Nicholas, David A. Ray, Michael Kube, Richard Reinhardt, Thomas H. Pringle, James Taylor, Russell C. Jones, Brett Nixon, Jean-Louis Dacheux, Hitoshi Niwa, Yoko Sekita, Xiaoqiu Huang, Alexander Stark, Pouya Kheradpour, Manolis Kellis, Paul Flicek, Yuan Chen, Caleb Webber, Ross Hardison, Joanne Nelson, Kym Hallsworth-Pepin, Kim Delehaunty, Chris Markovic, Pat Minx, Yucheng Feng, Colin Kremitzki, Makedonka Mitreva, Jarret Glasscock, Todd Wylie, Patricia Wohldmann, Prathapan Thiru, Michael N. Nhan, Craig S. Pohl, Scott M. Smith, Shunfeng Hou, Mikhail Nefedov[1], Pieter J. de Jong[1], Marilyn B. Renfree, Elaine R. Mardis & Richard K. Wilson

[1]Children's Hospital Oakland Research Institute, Bruce Lyon Research Building, 747 52nd Street, Oakland, California 94609, USA.

Nature 453, 175–183 (2008)

In this Article, Mikhail Nefedov and Pieter J. de Jong were omitted from the author list.

ERRATUM

doi:10.1038/nature07316

Tumour invasion and metastasis initiated by microRNA-10b in breast cancer

Li Ma, Julie Teruya-Feldstein & Robert A. Weinberg

Nature 449, 682–688 (2007)

In Fig. 4e of this Article, the two E-box sequences were inadvertently exchanged. E-box 1, which is near the stem-loop (at −313 bp), should be CACTTG instead of CACCTG, and E-box 2 (at −2,422 bp), which is distal to the stem-loop, should be CACCTG instead of CACTTG.

CORRIGENDUM

doi:10.1038/nature07253

Genome analysis of the platypus reveals unique signatures of evolution

Wesley C. Warren, LaDeana W. Hillier, Jennifer A. Marshall Graves, Ewan Birney, Chris P. Ponting, Frank Grützner, Katherine Belov, Webb Miller, Laura Clarke, Asif T. Chinwalla, Shiaw-Pyng Yang, Andreas Heger, Devin P. Locke, Pat Miethke, Paul D. Waters, Frédéric Veyrunes, Lucinda Fulton, Bob Fulton, Tina Graves, John Wallis, Xose S. Puente, Carlos López-Otín, Gonzalo R. Ordóñez, Evan E. Eichler, Lin Chen, Ze Cheng, Janine E. Deakin, Amber Alsop, Katherine Thompson, Patrick Kirby, Anthony T. Papenfuss, Matthew J. Wakefield, Tsviya Olender, Doron Lancet, Gavin A. Huttley, Arian F. A. Smit, Andrew Pask, Peter Temple-Smith, Mark A. Batzer, Jerilyn A. Walker, Miriam K. Konkel, Robert S. Harris, Camilla M. Whittington, Emily S. W. Wong, Neil J. Gemmell, Emmanuel Buschiazzo, Iris M. Vargas Jentzsch, Angelika Merkel, Juergen Schmitz, Anja Zemann, Gennady Churakov, Jan Ole Kriegs, Juergen Brosius, Elizabeth P. Murchison, Ravi Sachidanandam, Carly Smith, Gregory J. Hannon, Enkhjargal Tsend-Ayush, Daniel McMillan, Rosalind Attenborough, Willem Rens, Malcolm Ferguson-Smith, Christophe M. Lefèvre, Julie A. Sharp, Kevin R. Nicholas, David A. Ray, Michael Kube, Richard Reinhardt, Thomas H. Pringle, James Taylor, Russell C. Jones, Brett Nixon, Jean-Louis Dacheux, Hitoshi Niwa, Yoko Sekita, Xiaoqiu Huang, Alexander Stark, Pouya Kheradpour, Manolis Kellis, Paul Flicek, Yuan Chen, Caleb Webber, Ross Hardison, Joanne Nelson, Kym Hallsworth-Pepin, Kim Delehaunty, Chris Markovic, Pat Minx, Yucheng Feng, Colin Kremitzki, Makedonka Mitreva, Jarret Glasscock, Todd Wylie, Patricia Wohldmann, Prathapan Thiru, Michael N. Nhan, Craig S. Pohl, Scott M. Smith, Shunfeng Hou, Mikhail Nefedov[1], Pieter J. de Jong[1], Marilyn B. Renfree, Elaine R. Mardis & Richard K. Wilson

[1]Children's Hospital Oakland Research Institute, Bruce Lyon Research Building, 747 52nd Street, Oakland, California 94609, USA.

Nature 453, 175–183 (2008)

In this Article, Mikhail Nefedov and Pieter J. de Jong were omitted from the author list.

ERRATUM

doi:10.1038/nature07316

Tumour invasion and metastasis initiated by microRNA-10b in breast cancer

Li Ma, Julie Teruya-Feldstein & Robert A. Weinberg

Nature 449, 682–688 (2007)

In Fig. 4e of this Article, the two E-box sequences were inadvertently exchanged. E-box 1, which is near the stem-loop (at −313 bp), should be CACTTG instead of CACCTG, and E-box 2 (at −2,422 bp), which is distal to the stem-loop, should be CACCTG instead of CACTTG.

naturejobs

THE CAREERS MAGAZINE FOR SCIENTISTS

As the countdown to the switch on of the Large Hadron Collider (LHC) draws to a close, there is palpable excitement within the high-energy-physics community worldwide. The LHC, based at CERN, the European particle-physics laboratory near Geneva, will be the most powerful collider on the planet and is widely expected to deliver exciting new physics — as well as job opportunities.

But for those physicists working at Fermilab in Batavia, Illinois, any excitement about fresh results is tempered by the imminent demise of the Tevatron, the ageing accelerator that will shortly cede its crown to the LHC. It is doubly unfortunate, therefore, that US physics is also facing broad uncertainties over its federal budget.

Some suggest that the shift of the 'energy frontier' from the United States to Europe need not spell disaster for Fermilab, arguing that its stable of experts will be needed to help make sense of the steady stream of data emerging from the LHC (see page 258). After all, as Fermilab director, Pier Oddone, has pointed out, the lab's scientists are involved in both colliders. And for a little while, as the LHC reaches full operating status, there may be some healthy competition between the two as Fermilab makes a last-ditch attempt to detect the Higgs boson, the elusive particle that is one of the LHC's main goals.

But the experience of working on high-energy physics at Fermilab may begin to pall compared with what can be gained at the LHC, where the tacit knowledge associated with operating and tweaking the machine will feed discussions, elucidate problems and spark fresh insight. US-based researchers had hoped to be preparing to build and host the LHC's successor, the much-discussed International Linear Collider. But that has yet to be confirmed, and with the current budget woes for US physics, it seems unlikely to win approval. Money dedicated to projects one year can disappear the next on Congress's whim, which makes hosting huge international projects difficult. Those interested in a career in high-energy physics may therefore continue to head towards Europe for some years to come.

Gene Russo is editor of *Naturejobs*.

CONTACTS
Editor: Gene Russo

European Head Office, London
The Macmillan Building,
4 Crinan Street, London N1 9XW, UK
Tel: +44 (0) 20 7843 4961
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

European Sales Manager:
Andy Douglas (4975)
e-mail: a.douglas@nature.com
Business Development Manager:
Amelie Pequignot (4974)
e-mail: a.pequignot@nature.com
***Nature*events:**
Claudia Paulsen Young (+44 (0) 20 7014 4015)
e-mail: c.paulsenyoung@nature.com
France/Switzerland/Belgium:
Muriel Lestringuez (4994)
Southwest UK/RoW: Nils Moeller (4953)

Scandinavia/Spain/Portugal/Italy:
Evelina Rubio-Hakansson (4973)
Northeast UK/Ireland:
Matthew Ward (+44 (0) 20 7014 4059)
North Germany/The Netherlands:
Reya Silao (4970)
South Germany/Austria:
Hildi Rowland (+44 (0) 20 7014 4084)

Advertising Production Manager:
Stephen Russell
To send materials use London address above.
Tel: +44 (0) 20 7843 4816
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com
***Naturejobs* web development:** Tom Hancock
***Naturejobs* online production:** Dennis Chu

US Head Office, New York
75 Varick Street, 9th Floor,
New York, NY 10013-1917
Tel: +1 800 989 7718
Fax: +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

India
Vikas Chawla (+91 1242881057)
e-mail: v.chawla@nature.com

Japan Head Office, Tokyo
Chiyoda Building, 2-37 Ichigayatamachi,
Shinjuku-ku, Tokyo 162-0843
Tel: +81 3 3267 8751
Fax: +81 3 3267 8746

Asia-Pacific Sales Manager:
Ayako Watanabe (+81 3 3267 8765)
e-mail: a.watanabe@natureasia.com
Business Development Manager, Greater China/Singapore:
Gloria To (+852 2811 7191)
e-mail: g.to@natureasia.com

FERMILAB

Collision course

This month, all eyes in the high-energy-physics community will be on the long-awaited launch of CERN's new particle collider. But US budget cuts and an uncertain job market mean the field has little else to celebrate. **Eric Hand** reports.

Dave Schmitz spent the early part of last year's Christmas holiday at home in Chicago, poring over his graduate thesis and postdoc fellowship application essays. It was a big career step for Schmitz, a neutrino physicist about to finish his PhD at Columbia University in New York. He was writing a five-page research statement, part of the application for the Lederman Fellowship, a prestigious postdoc position at the Fermi National Accelerator Laboratory in Batavia, Illinois.

Then the budget news broke. On 19 December, the US Congress passed a year-end spending bill that devastated the high-energy-physics community, reducing its budget by $94 million (see *Nature* **451**, 2–3; 2007). Within days, Fermilab announced that it would lay off around 200 employees — about 10% of its staff. It also mandated that all staff take a week of unpaid leave every other month. Stunned, Schmitz e-mailed the head of the fellowship selection committee to ask her if he should even bother applying. "Do I want to start my career at a place that's having such difficulty?" he recalls thinking. "It gave me serious pause, and made me think that maybe I should consider something else."

The budget woes were just the beginning for the US physics community. Throughout 2008, signs of recession deepened, casting a pall over lucrative non-academic physics jobs in technology or on Wall Street, a job market that tends to march hand in hand with the economy. And all of this on the eve of a once-in-a-decade event in high-energy physics: the opening of a new collider. This month, the Large Hadron Collider (LHC) at CERN, Europe's particle accelerator centre near Geneva, Switzerland, will start smashing particles together at the highest energies the world has ever seen. Fermilab's Tevatron will be eclipsed as the world's most powerful accelerator, putting the lab at a further disadvantage as a destination for top high-energy-physics talent.

F. ULLRICH/FERMILAB

"You cannot expect to go on for a decade without capital investment and still have competitive facilities."
— Pier Oddone

With fewer than 300 research scientists and little more than 50 postdocs, Fermilab employs only a small proportion of the high-energy-physics community. But thousands of university scientists — most of them in the United States — are affiliated users. And, as the only dedicated particle-physics laboratory in the United States, Fermilab is a bellwether for the community, a place that is watched closely. Right now the outlook is fairly bleak.

When the Tevatron is shut down in 2009 or 2010, the United States will not have laboratories exploring the energy frontier of particle physics for the first time since the 1930s, when the early accelerators were built. And little relief is expected, says Fermilab director Pier Oddone, who laments the lack of investment in other experiments. Much time, effort and money were put into planning the International Linear Collider (ILC), a successor to the LHC that Fermilab hoped to build in Illinois. The future of the ILC is now tenuous at best. "You cannot expect to go on for a decade without capital investment and still have competitive facilities," Oddone says. "There's a real crunch right now."

For now, though, Schmitz and others have a reprieve. Schmitz got the fellowship. A few weeks before he began in June, an anonymous donor gave Fermilab's operators $5 million to avert the unpaid furloughs. And several weeks after that, Congress tacked on $62.5 million for high-energy physics to a supplemental

spending bill for the war in Iraq, putting an end to lay-offs.

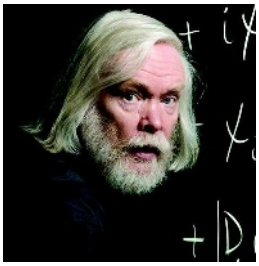
Although the immediate crisis has been averted, the long-term health of high-energy physics remains in question, says Roman Czujko, director of the Statistical Research Center of the American Institute of Physics (AIP) in College Park, Maryland. "The decline in the United States — this is the big deal," he says. In fact, that researchers can easily get a postdoc position might be a barometer for how bad things are, says Czujko. "The postdoc numbers invariably go up during a downturn in the economy," he says. "Instead of buying a piece of equipment, you buy yourself a postdoc for a year."

A rise in postdoc numbers during the past year could also reflect a downturn in entry-level non-academic industry jobs, leading physicists to opt to continue their studies a little longer. But in the long term, the opening of the LHC, alongside the decline of accelerators at Fermilab, could put the squeeze on academic physics postdoc positions in the United States. Czujko says this is especially relevant to high-energy physics. According to a 2007 AIP survey, almost four out of five students with PhDs in high-energy physics go on to postdocs rather than potentially permanent jobs in the workforce. That's a higher proportion than in most of the more practical subcategories of awarded physics PhDs, such as applied physics, where more than 50% are able to secure potentially permanent jobs.

As a result, Czujko says, high-energy physicists are more susceptible to the vagaries of federal funding of national research laboratories such as Fermilab. Eventually, he expects the number of high-energy-physics PhDs and postdocs to decline. "It will be tougher for them to find support in the United States. Those who do — they'll have to avail themselves of the equipment at CERN."

CERN and the LHC are already a big draw. As a string theorist, Fermilab postdoc Mark Jackson doesn't need to be near CERN's equipment. But he still wants to be in Europe to be near the heart of the action. Currently in the final year of his postdoc, Jackson has just accepted a position at Leiden University in the Netherlands. Like five of his colleagues, he'll be following the new machine.

At Fermilab, his postdoc pays about $50,000.



R. HAHN/FERMILAB

CERN

Fermilab Lederman fellows Dave Schmitz (top) and Diego Tonelli (centre), and CERN theorist John Ellis.

Leiden University will pay a base salary of €39,000 (US$57,000). But it wasn't a little extra money that seduced him. "Anyone doing physics isn't too concerned about their salaries," he says. The draw was quick and ready access to the data soon to pour out of the LHC, as well as Planck, a European Space Agency mission scheduled to launch in early 2009 that will study the 'cosmic microwave background' — that is, radiation from the Big Bang.

Fermilab is fighting to stay relevant and remain a draw for top talent — partly through LHC collaborations. At the Remote Operations Center, Fermilab employees can work on one of the LHC's two main detector experiments, which are connected by a high-throughput computer-networking grid. That, combined with Fermilab's stable of theorists and technicians who already familiar with the trials and tribulations of the Tevatron, will make Fermilab a fertile crucible for good work, says Oddone. "What we have aimed for here is to try to have a critical mass, such that, once the detectors are running at the LHC, the experience of coming to Fermilab to do physics is as rich as it would be to go to CERN," he says.

Others aren't convinced. Steve Nahn, an assistant professor at the Massachusetts Institute of Technology in Cambridge, is sceptical that the remote centre will really mimic the experience of working at the LHC. "There's no crucial operation that it does," he says. "If it were unmanned, the show would go on. And that will be its problem." He adds that being an ocean away means missing out on the hallway conversations or discussions in an expert's office that elucidate the day-to-day quirks of operating one of the most complicated machines on Earth. He has sent all four of his graduate students and postdocs to Europe to work full-time at CERN.

Many US universities are following suit. Of the 9,300 outside 'users' of CERN — a number that is up 50% since 2004 — some 1,400 are American, according to John Ellis, a senior theorist at CERN. "They are now the second-largest national contingent here, exceeded only slightly by Italy, and way ahead of Germany, France and the United Kingdom," he says.

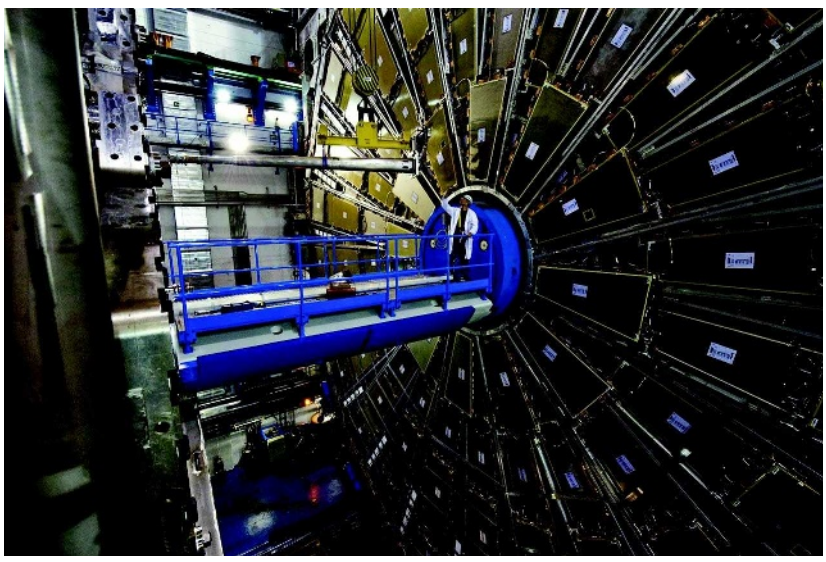
But as at Fermilab, the number of staff physicist positions are few and hard to come by. In fact, CERN is having to cut back on its staff PhD physicists, says James Purvis, CERN's head of recruitment. During the past decade the number has dropped from 100 to 78, he says. "The only way of building the [LHC] project with a constant budget was with fewer people."

The high-energy-physics job market at universities doesn't seem to be faring much better. In the United States, retirement and vacancy rates for physics faculty members are going down as the number of temporary or non-tenure track physics faculty goes up, according to AIP surveys. "There are certainly many more people than there are spots," says Nahn, who got his tenure-track job at MIT only after six years as a postdoc. "The competition is pretty steep."

Many hope the LHC will not only generate new findings, but ample high-energy-physics research and career opportunities. "If this machine at CERN does even half the things we expect it do, it will be great," says Diego Tonelli, an Italian who is in the second year of a Lederman fellowship at Fermilab. "We'll have another 50 years of nice times." ■

Eric Hand is a *Nature* correspondent based in Washington DC.

M. BRICE/CERN



CERN's Large Hadron Collider will be the world's highest-energy particle accelerator.

NATURE|Vol 455|11 September 2008

MOVERS

Eric Barron, director, National Center for Atmospheric Research, Boulder, Colorado

2006–08: Dean, Jackson School of Geosciences, University of Texas at Austin
2002–06: Dean, College of Earth and Mineral Sciences, Pennsylvania State University, University Park, Pennsylvania
1989–2006: Professor of geosciences, Pennsylvania State University, University Park, Pennsylvania

Eric Barron has had a controversial start as the new director of the National Center for Atmospheric Research (NCAR) in Boulder, Colorado. He had looked forward to leading the place that had been the source of a fellowship and multiple collaborations over the years. And yet, soon after taking up his new post in July, Barron closed NCAR's Center for Capacity Building and laid off its director, Mickey Glantz, a climatologist who helps developing countries deal with climate change (see *Nature* 454, 808–809; 2008). Barron blames stagnant budgets, which have forced NCAR to lay off 12% of its core staff over the past five years.

It is a sharp contrast with his previous job. For two years, Barron ran the Jackson School of Geosciences at the University of Texas, founded in 2002 with a $237-million gift from John and Katherine Jackson — at the time, the largest ever to a US public university. The Jackson money allowed the university to expand rapidly; Barron hired several people, including seven new faculty members in climate science.

But Barron had an affinity for NCAR. After studying geology as an undergraduate and oceanography as a PhD student in Florida, he received a Cray supercomputing graduate fellowship from NCAR in 1976 and turned his attention to climate modelling. Soon he was serving as editor at various geosciences journals and as panellist on numerous advisory committees.

Through nearly two decades of working at Pennsylvania State University, Barron maintained strong links with NCAR, collaborating with researchers and serving on the board of trustees for the University Corporation for Atmospheric Research, which manages NCAR. That history is what led him back to Boulder, he says — in particular, the notion of serving society at large through NCAR's work.

"It's an extraordinary group of people doing something of considerable importance for society," he says. And even as finances drive him to lay off some of those people, he is planning to find budget-friendly ways to bolster the centre's impact — for instance, by developing partnerships with outside institutions to share NCAR expertise.

John Dutton, the now-retired former dean of the college of Earth and mineral sciences at Pennsylvania State University, lauds his one-time hire for engaging students and helping them achieve their academic aims. "He has a very positive spirit and gets things done," says Dutton, adding that Barron has the management style to cope with NCAR's budget woes. ■

Alexandra Witze

NETWORKS & SUPPORT

Fellowships at the FDA

The US Food and Drug Administration (FDA) hopes that a newly established annual fellowship programme will help reel in a diverse mix of top-notch PhDs, PharmDs, DVMs and MDs. "We're trying to recruit a broad mix of the brightest researchers and clinicians, with the fellowship as a way to recruit future FDA employees," says Frank Torti, the agency's new chief scientist who fast-tracked the programme's creation (see *Nature* **453,** 560; 2008). So far, interest in the programme, a two-year stint combining coursework and research, has been strong. By its 30 August deadline, the agency had received more than 600 applications for its 40 positions.

In addition to MDs and clinicians, Torti wants to recruit pharmaceutical professionals, epidemiologists, statisticians and psychologists to tackle new scientific areas at the agency. Whether they have FDA, industry or academic aspirations, Torti says the FDA fellows will be hot commodities — especially in pioneering research areas such as regenerative medicine. Torti expects six to ten will specialize in biomedical engineering. "With so many novel devices coming through, it is a booming area of regulation," he says.

As the FDA regulates 25% of the manufactured goods that make up the US gross domestic product, it connects fellows with a wide array of industries seeking candidates who understand the inner workings of the FDA. Academia may seem an odd beneficiary, but Torti says people who understand FDA regulatory processes will be more effective in designing trials to meet FDA requirements.

Fellows will be chosen by a 'mutual selection' process. Eligible applicants will choose from 100–120 preceptors working in the various FDA centres, which specialize in the evaluation of drugs, biologics, devices, food safety, nutrition and veterinary medicine. The preceptors will then rank those fellows. Sanjai Kumar, a preceptor and chief of the malaria research programme at the Center for Biologics Evaluation and Research, says he will look for applicants who express a clear desire to forge a career in the regulatory sciences. Most fellows will be housed at the White Oak facility in Silver Spring, Maryland, but some will take courses electronically at the National Center for Toxicologic Research near Little Rock, Arkansas.

First-year courses will cover FDA law, policy, management, trial design, epidemiology and statistics. Over 70% of time will be devoted to a hypothesis-driven research project. ■

Virginia Gewin

POSTDOC JOURNAL

Job transplant

I'm uprooting myself, and uprooting requires digging. This month I've made academic discoveries not worthy of publication, such as the pile of important papers I put aside to read at the start of my postdoc, or the helpfully underlined answers in a page of neat algebra (if only I still had the questions). These will not be accompanying me to my new position at University College Dublin.

It's undesirable but unavoidable. Uprooting causes root damage. Once more I'm removing myself from a network of friends and familiar surroundings. I'm going to be replacing tangible, face-to-face collaborations with remote working relationships and all the problems they bring. The advice of trusted colleagues will no longer be down the corridor, and my ex-student cannot stroll down the hall to consult me. I am making myself peripheral to my current work circles.

Uprooting is followed, of course, by replanting in a carefully chosen, fertile location. I'm taking a new post in a hitherto undeveloped area for my new department. It's exciting and challenging. I'll need to develop joint projects and nurture connections with other research groups. And all this in the luxury of my native tongue, making communication a little bit easier and my partner's job prospects more promising. With a bit more careful digging, her roots will be ready to join mine. I think the prognosis for this transplant is good. ■

Jon Yearsley is a senior postdoc in evolutionary genetics at the University of Lausanne in Switzerland.

NATUREJOBS is pleased to present a HIGHLIGHT ON **IRELAND**

***NATURE* ISSUE: 2nd OCTOBER 2008**

DEADLINE FOR ADVERTISERS:
26th September 2008

The Highlight on Ireland is exclusively dedicated to promoting scientific jobs in Ireland; with *Nature*'s worldwide circulation of over 56,000*, this highlight will provide significant exposure.

Ireland's success in research and development is widespread across its businesses and sectors. This dynamic research environment is fast growing with the support of the Government's investment of €8.2 billion for seven years announced in 2006. Government departments, academia, funding agencies and regulatory authorities are interconnected and all contribute to increasing Ireland's research capabilities into world class.**

This Highlight will be a valuable reference for researchers, students and educators in search for career opportunities and discipline-related events in Ireland. It will be eagerly read by those drawn to the specially marked section.

All job advertisements will receive a complimentary 8-week online posting on *naturejobs*.com – the largest dedicated job board for the scientific community with over 5,000*** jobs!

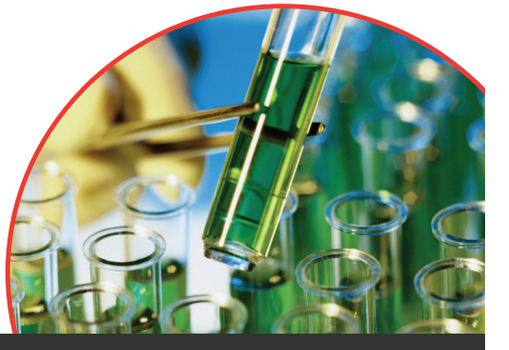
Contact your representative to take part in this opportunity today:

Matt Ward
T. +44 (0)20 7014 4059
F. +44 (0)20 7843 4996
E. m.ward@nature.com

IN169945R

If you are interested in advertising events, please contact Ghizlaine Ababou
+44 (0)20 7014 4015, g.ababou@nature.com

*Publisher's Data 2008
** www.idaireland.com, 2008
***July 2008



WWW.NATUREJOBS.COM

naturejobs

nature publishing group npg

The whaleblimp herder

Transport issues.

Chris Butler

The dog crouched down while the whaleblimp floated over it with a few inches of clearance, then scampered after it giving playful chase. Jones smiled and swung his crook through the grass, sending dandelion heads tumbling in the breeze.

In the distance he could see something, a cloud of dust kicked up from the land. He frowned, crouched down and pressed his hand to the ground. Men on horseback coming, he decided.

The sun was high in the sky so he kept in the shade of one of the whaleblimps. Crickets chirruped all around, but their sound was lost in the clatter of hooves and the rattle of a cart as the horsemen arrived. One of the whaleblimps was spooked and drifted off.

"Patch," Jones called to the dog. "Quit playing and start working."

Jones blew into his shepherd pipes and Patch got the idea. The dog bounded away and barked at the wayward blimp, bringing it back into the herd.

"Hey kid," one of the men said, and climbed down. He was little more than a kid himself, Jones thought, but he had the swagger of someone who fancied himself the leader of his group. "Name's River. Where're you going with these?"

Jones shifted the backpack on his aching shoulders and flapped away a fly buzzing round his face. The herd had slowed to a halt. He squinted against the sun and tried to figure how much these men might know about whaleblimps.

"I'm taking food and supplies to the people at the Moor power plant," he said.

River eyed the panniers suspended beneath the blimps. He nodded to his men and they pulled on the rope to drag one down, then hauled back the tarpaulin to inspect the cargo.

"You're lucky to have the power plant," Jones said. "I hear they couldn't keep York going any more."

River nodded. "Yeah, I heard that too. But see, you shouldn't just be herding these things across my land without asking first. There are tolls to be paid."

"I'm doing this for you," Jones said. "Without power …"

River spat into the dirt. "Yeah, well, we need power but we also need to eat."

To Jones, the three men did not look like they were starving. The horses looked scrawny, though.

"My men are going to unload one of those panniers."

Jones sighed and nodded. He was getting used to this kind of trouble. Times were hard wherever he went. Not for the first time he thought it was a good thing the whale bioforms were inedible. His trade would have ended long before now.

"And if you want to take these back across my land after you're done," River said, "you better be able to pay me then, too."

"But …"

River struck Jones with the back of his gloved hand, knocking him down. Patch barked angrily and ran forward.

"Quiet down, Patch," Jones said. He could let one knockdown go for now. He'd had worse before and it was not enough to deter him from what he had to do.

"Times are tough, kid," River said. "I'll take a fifth of whatever they're paying you. And I'll know what they've paid you, so don't try lying about that."

Jones figured that was a bluff. He watched while the men transferred the contents of the pannier onto their cart.

"I'll find you on your way back out of town," River said. "You shouldn't be too hard to spot."

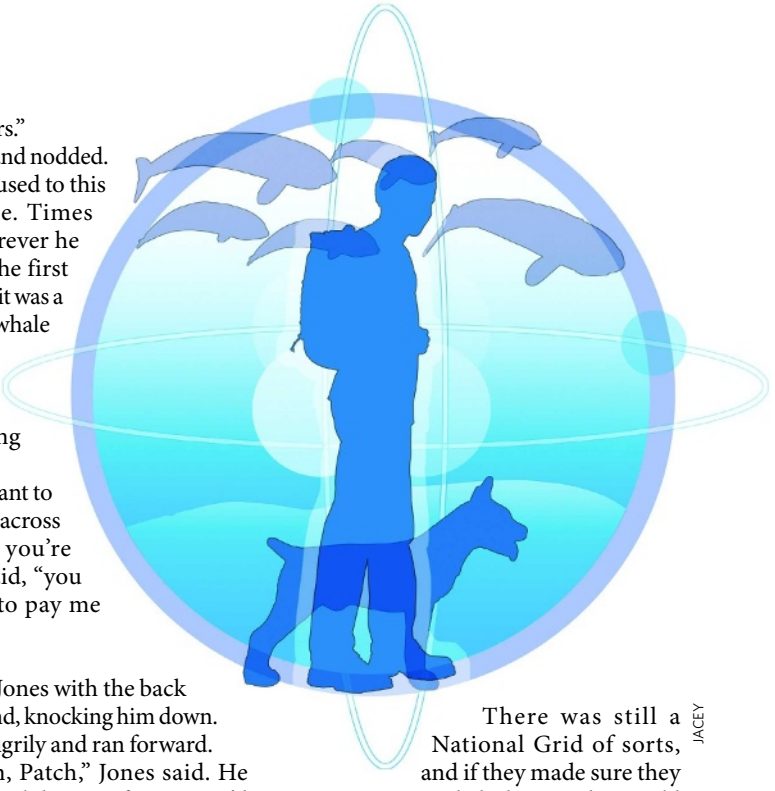
A herd of airborne blimps could be seen for miles, and they moved slow. The three men rode away, hollering to each other and to the sky.

When Jones reached the power plant he pulled the papers from his backpack and showed them at the gate. "I'll need a word with your security people," he said.

"Sure," the guard said, "I'll have someone come down to talk to you. Meanwhile, if you'd like to take your herd through the gate, I'll call someone to hook up the blimps for you."

"Thanks."

Jones and Patch drove the herd inside. The supplies were unloaded and carried away. Then came the real purpose of the delivery. Jones watched as a team of men drained off the helium transported in the blimps. The gas had been taken out of the system at York. It would keep this plant's cooling system going for a while longer. York's loss was their gain.



JACEY

There was still a National Grid of sorts, and if they made sure they recycled whatever they could from the stations they closed, they could keep others going for a while yet.

Jones was loading his payment into his backpack when the head of security arrived to talk to him; the traded medical supplies were needed desperately back home at Harrogate.

"I had a guy named River making life difficult for me on the way over here," Jones said. "I'm not doing business with you again unless you keep control of the locals."

"People are getting desperate," he said, but he apologized. "We wouldn't want to lose a good courier like you."

Jones didn't know where his next helium cargo would take him. Maybe here. Maybe someplace else. "I don't want to see him again," he said.

A man came in with the blimps, vacuum sealed for the journey home. Jones put them in his backpack too. It was the easiest and safest way to transport them.

At the exit gate the guard said: "Hardly any Moon tonight."

Jones already knew the phase of the Moon, and left nothing to chance.

He pulled on his night vision goggles, whistled to Patch and set off on the long walk back. ■

Chris Butler's stories have appeared in *Asimov's Science Fiction* magazine and *Interzone*. His novel, *Any Time Now*, was published by Wildside Press. Latest news can be found at www.chris-butler.co.uk.

FUTURES